

QUANTA 2006

Protein Design

MAY 2006

Copyright

(1) Copyright

Copyright ©2006, Accelrys Software Inc. All rights reserved. The Accelrys® name and logo are registered trademarks of Accelrys Software Inc.

This product (software and/or documentation) is furnished under a License/Purchase Agreement and may be used only in accordance with the terms of such agreement.

(2) Trademark

The registered trademarks or trademarks of Accelrys Software Inc. include but are not limited to: ACCELRYSS® & ACCELRYSS Logo, ACCORD, BIOSYM®, CATALYST®, CERIOUS®, CERIOUS2®, CHARMM®, CHEMEXPLORER®, DIAMOND DISCOVERY®, DISCOVER®, DISCOVERY STUDIO®, DIVA®, FLEXSERVICES®, GCG®, GENEATLAS®, INSIGHT®, INSIGHT II®, MACVECTOR®, MATERIALS STUDIO®, OMIGA®, QUANTA®, SEQARRAY, SEQFOLD®, SEQLAB®, SEQMERGE®, SEQSTORE®, SEQWEB®, TOPKAT®, TSAR®, UNICHEM®, WEBKIT, WEBLAB®, WISCONSIN PACKAGE®. All other trademarks are the property of their respective holders.

(3) Restrictions on Government Use

This is a “commercial” product. Use, release, duplication, or disclosure by the United States Government agencies is subject to restrictions set forth in DFARS 252.227-7013 or FAR 52.227-19, as applicable, and any successor rules and regulations.

(4) Acknowledgments, and References

To print photographs or files of computational results (figures and/or data) obtained using Accelrys software, acknowledge the source in an appropriate format. For example:

“Computational results obtained using software programs from Accelrys Software Inc. Dynamics calculations performed with the Discover program using the CFF forcefield, ab initio calculations performed with the DMol³ program, and graphical displays generated with the Cerius² molecular modeling system.”

To reference an Accelrys publication in another publication, Accelrys Software, Inc., is the author and the publisher. For example:

Accelrys Software Inc., Cerius² Modeling Environment, Release 4.8, San Diego: Accelrys Software Inc., 2006.

(5) Request for Permission to Reprint and Acknowledgment

Accelrys may grant permission to republish or reprint its copyrighted materials. Requests should be submitted to Accelrys Scientific Support, either through electronic mail to support@accelrys.com, or in writing to:

Accelrys Scientific Support
10188 Telesis Court, Suite 100
San Diego, CA 92121

Please include an acknowledgement “Reprinted with permission from Accelrys Software Inc., Document name, Month Year, Accelrys Software Inc., San Diego.” For example:

Reprinted with permission from Accelrys Software Inc., Cerius² User Guide, June 2006, Accelrys Software Inc.: San Diego.

Contents

Preface: Protein Design	1
Overviewing the Protein Design palette	2
Protein MODELER	3
1. The Sequence Viewer	5
Overview	5
Sequence Data	5
Saving Sequences And Alignment Between Sessions	5
Changing Maximum Number of Sequences	5
The Sequence Viewer	6
Display of Graphs	7
The Sequence Viewer icons	7
2. Reading and Writing Sequence Data Files	11
Overview	11
Reading and Writing Sequence Data Files	11
Read Sequence/Alignment File	11
Read Sequence Data File	12
Demo of User Data	12
To run the demo:	13
Reference	14
Write Sequence File	14
Plot Sequence Viewer	14
Remove Sequence	15
3. Protein Utilities	17
Overview	17
Simple Representations of Proteins	17
Tools and Options	18
Color by Structure Properties	19
Color By Sequence Properties	20
Color by Homology	20
4. Protein Editor	21
Overview	21
Editing Proteins	21
Ideal Residue Definitions	21
Regularization	21
Residues	22
Editing Segments	22
Hydrogen Addition	22
Tools and Options	23
Amino Acid Selection	23
The Editing Tools	23
4. Predict Secondary Structure	27
Overview	27
Predicting Secondary Structures	27
Momany Prediction	28
Holley/Karplus Prediction	28
GOR Prediction	29

Conservation Profiles	29
Hydrophobicity Scales	29
Sequence Viewer Plots	30
Saving Predictions	31
Tools	31
5. Align and Superpose	33
Overview	33
Aligning and Superposing Sequences	33
Using Active Sequences and Active Ranges	33
Criteria for Aligning and Matching Sequences	34
Alignment	35
Manual Alignment Editing	37
Saving and Restoring Alignments	37
Dot Plots	38
Alignment Constraints	39
Matching Residues	39
Color by Homology	40
Superposing Structures	40
Tools and Options	41
The Constraints Palette	45
Constraint Palette Tools	45
6. Create Homology Model	47
Overview	47
Copying Homology	47
Tools	48
7. Superpose Folding Motif	51
Overview	51
Superposing Folding Motifs	52
Protein Geometry	53
Secondary Structure Representation	54
Sequence Alignment	54
Tools and Options	54
Select Active Secondary Structure	55
Match Secondary Structure	55
Reviewing the Matches	58
Superpose and Align Molecules	58
Demonstration of Using Superpose Motif	59
8. Model Backbone	61
Overview	61
Modeling the Protein Backbone	61
Regularizing Regions	62
Folding Residues	62
Fragment Searching	63
Tools and Options	64
9. Model Side Chains	69
Overview	69
Modeling Sidechains	69
Close Contacts	70
Rotamers	70
Spinning Side Chains	71
Tools and Options	71
10. Analyze Secondary Structure	75

Overview	75
Analyzing Secondary Structures	75
Hydrogen Bond Calculations.	75
Secondary Structure Assignment	76
Tools and Options	77
11. Calculate Accessibility	79
Overview	79
Reference.	79
Calculating Accessibility	79
Accessibility Calculations	80
Contact Area	81
Displaying Accessibility and Contact Areas	81
Tools and Options	82
12. Display Contact Maps	85
Overview	85
Calculating Contact Maps.	85
Plotting Method	86
Secondary Structure Elements	87
Molecule Display	88
Difference Contact Maps.	88
Distance Contact Maps	88
Energy Contact Maps	89
Interaction-Type Contact Maps	89
Hydrogen Bonds	90
Residue Type	90
Tools and Options	90
13. Analyze Domain Structure	93
Overview	93
Analyzing Domain Structures.	93
The Clustering Algorithm	94
Loop Regions	94
Tools and Options	95
14. Profile Analysis	99
Overview	99
Analyzing Protein Profiles	99
Comparing a Profile to a Sequence.	100
Plotting Profiles.	100
Comparing Profiles to Other Sequences	100
Tools and Options	101
15. Protein Information	103
Overview	103
Retrieving Protein Information.	103
Tools.	104
Running a Protein Information Query	105
16. Sequence Database	107
Overview	107
FASTA Sequence Searching	107
Tools.	108
17. Structural Database	109
Overview	109
Searching the Structure Database.	110
Defining A Structure Database Query	110

Tools and Options	113
18. Motif Database	121
Overview	121
Searching the Motif Database	121
Tools and Options	122
Motif Database Log File	124
19. Protein Health	127
Overview	127
Using Protein Health	127
Analysis of Multiple Conformations	128
Tools and Options	130
Dunbrack and Karplus rotamer definitions	131
20. Using MODELER	137
Overview	137
MODELER	138
Accessing MODELER	139
Displaying MODELER results	141
Evaluating MODELER Results	141
Deleting files after a MODELER run	143
MODELER files and file modifications	143
The MODELER control file	144
Modifying files to modify models	146
Defining MODELER Restraints	147
MODELER Theory	148
Sequence-structure alignment (Align2D)	150
Structure alignment (Align3D)	152
Building the structure	153
Probability density functions	154
Probability density functions observed in known protein structures	154
Probability density functions for bond lengths, bond angles, and dihedrals	155
Atom-atom repulsions	157
Distance between two C α atoms	157
Distance between main-chain N and O atoms	160
Distances between sidechain-sidechain and sidechain- mainchain atoms	160
Main-chain conformation	160
Side-chain conformation restraints	162
Restraining Chi ₁ side-chain dihedral angles	163
Restraining Chi ₂ dihedral angles	163
Restraining Chi ₃ dihedral angles	163
Restraining Chi ₄ dihedral angles	164
Basis and feature probability density functions	165
Feature PDFs used for restraining model protein features	166
C α -C α distance feature PDF	166
Main-chain N-O distance feature PDF	167
Stereochemical feature PDF	167
Main-chain conformation feature PDFs	167
Side-chain dihedral angle feature PDFs	168
Molecular PDF used for structure generation	168
Structure generation using MODELER	169

Modeling loops	170
References	172
A. Conversion of External Sequence Data Files to QUANTA Format	175
Sequence Data File Format.	175
Sequence User Color File Format	176
B. Creating a Fragment dmfile	179
C. Customizing the Databases	181
Overview	181
The PDB Master File	181
Running CREBASE to create the Database Files.	182
Creating the MSF Library.	183
D. The Geometric Structure Definition File	185
The file format.	186
E. The Protein Parameter File	189
F. Read Sequence File Formats	193
Overview	193
Pearson (FASTA) format (extension .aa)	193
GCG (extension .gcg).	193
HAHU	194
NBRF-PIR	194
SWISSPROT (extension .sws).	195
G. Running the Search Standalone	197
Overview	197
Search Commands	197
Running a Search	199
H. Torsion Angles and Centers	201
Overview	201
Database file format	202
I. Wildcard Residue Type File	203
J. MolScript	205

Protein Design is a versatile application for modeling and analyzing protein structures. There are two major palettes associated with this application, Protein Design and Protein Utilities. The Protein Design palette contains 18 options that can be classified into three types of utilities. The Protein Utilities palette is also displayed and contains visual tools and structure checks, using the Protein Health option.

- Modeling
 - Align and Superpose**
 - Create Homology Model**
 - Edit Protein**
 - Model Backbone**
 - Model Side Chains**
 - Predict Secondary Structure**
 - Superpose Folding Motif**
- Analysis
 - Analyze Domain Structure**
 - Analyze Secondary Structure**
 - Calculate Accessibility**
 - Display Contact Maps**
 - Profile Analysis**
- Databases
 - Motif Database**
 - Protein Information**
 - Sequence Database**
 - Structure Database**

This reference book is designed to give a general description of each of the utility interfaces listed above including the scientific methods, and options and tools. The information is arranged in alphabetical order by palettes.

Overviewing the Protein Design palette

<i>Align and Superpose</i>	A variety of options are available for aligning sequences, identifying homologous regions, and superpositioning structures.
<i>Analyze Domain Structure</i>	Protein Design uses geometric relations between secondary structural elements to automatically identify domains, and provides tools that allow you to define and edit the domains.
<i>Analyze Secondary Structure</i>	This utility provides tools that assign secondary structures to proteins. By defining the secondary structure of a molecule, the shape of the molecule can be visualized. This is a two-step procedure, wherein hydrogen bonds are calculated, then secondary structures are assigned based on those hydrogen bond patterns and phi/psi angles.
<i>Calculate Accessibility</i>	The tools in this utility determine the solvent accessibility of a molecule or the contact area between molecules or regions of molecules.
<i>Create Homology Model</i>	This utility enables you to copy coordinates from a known structure or structures to a sequence whose structure is unknown.
<i>Display Contact Maps</i>	The most common use of contact maps is to show the inter-residue $C\alpha$ - $C\alpha$ distances. Similar properties which are a function of two residues could also be plotted in similar fashion, such as inter-residue VDW energy or number of inter-residue hydrogen bonds. The properties currently displayed as contact maps in this utility are: $C\alpha$ - $C\alpha$ distances; $C\beta$ - $C\beta$ and side chain contact distances; van der Waals interaction energy; electrostatic interaction energy; total interaction energy; hydrogen bonds; and residue type interactions.
<i>Edit Protein</i>	The protein editor provides tools to modify the sequence of a protein by mutating, inserting or deleting residues. There are also tools to enter a new sequence, generate an MSF, or change the hydrogen representation of the molecule.
<i>Model Backbone</i>	The protein modeling tools are divided into two utilities: Model Backbone for defining main chain conformations, and Model Side Chains for defining sidechain conformations. They are closely interdependent. This utility includes tools to search the structure database for fragments to use in model building and a regularization and energy minimization tool.
<i>Model Side Chains</i>	This utility contains tools for modeling the protein side chain conformations. It is assumed that the protein main chain has been determined using the Model Backbone utility. This utility includes rotamer libraries and regularization and energy minimization.

<i>Motif Database</i>	This module provides tools that search a database for structures with similar folds to one active molecule. The search motif can be entire proteins or selected substructures.
<i>Predict Secondary Structure</i>	This utility is concerned with sequence analysis. It can be used for sequences for which the structure coordinates are not defined. There are five different types of analysis that can be displayed, three of which are prediction methods. The results of the analysis are usually presented as plots of property vs. sequence position.
<i>Profile Analysis</i>	This utility follows the method of Bowie, Luthy, and Eisenberg. 3D protein structures are analyzed into 1D profile sequences. This method is used to generate a plot of the quality of a model.
<i>Protein Information</i>	This utility retrieves textual information on PDB files from the protein structure database by accessing the QUANTA file \$HYD_LIB/database.dat. This database file contains information on all the PDB files currently in the Brookhaven Protein Databank. It is also the same data file used by the structural database utility.
<i>Sequence Database</i>	This utility provides a group of options to search the Protein Data Bank for sequences that closely match a specific sequence. QUANTA uses the FASTA sequence search algorithm.
<i>Structural Database</i>	Searching the structural database aids in molecular modeling. Using this utility, a search can be performed on a specified sequence or conformation against all of the known protein structures from the Brookhaven Protein Databank. This information is found in the file \$HYD_LIB/database.dat.
<i>Superpose Folding Motif</i>	This utility superposes structures on the basis of their overall folding rather than requiring identifying homologous residues. Using this utility, protein structures with similar folding motifs, but possibly little other obvious homology, can be superposed.

Protein MODELER

The Protein MODELER application provides an interface to the automated homology modelling program MODELER. The application includes the **Align and Superpose** utility identical to that in Protein Design and tools to run and read MODELER results. There are tools for display and analysis of MODELER results.

1. The Sequence Viewer

Overview

The Sequence Viewer window comes up automatically upon entering Protein Design, Protein MODELER, Protein Health or Protein Profile Analysis and closes when exiting these applications. The window can be iconized or expanded by clicking the appropriate icons in the top right of the frame of the window.

This chapter describes:

- The Sequence Viewer
- Display of Graphs
- The Sequence Viewer icons

Sequence Data

Sequences, for which the only information is the amino acid sequence, are supported and can be read in using the utilities in the **Sequence Data** option under the **Files** pulldown. Sequences can be converted to MSFs using the tool in the Protein Editor utility.

Saving Sequences And Alignment Between Sessions

While the program is running, a file called protein_default.aln (which is in an extension of Clustal format) is kept up to date with the current sequence selection and alignment. On exiting QUANTA, this file is written to the constants file, .cst, to be saved until the next session.

Changing Maximum Number of Sequences

By default, Protein Design handles up to 30 sequences and 50,000 residues in all the sequences. There is also a limit of 2000 columns in the sequence viewer. These values can be reset by typing the command SEQM into the command line. You will get a dialog box in which you can enter the required maximum dimensions. These values will be saved and used in future QUANTA sessions.

The Sequence Viewer

Using the mouse

In general, clicking any item on the Sequence Viewer is done with the left mouse button except for those operations which drag either a sequence or slider and these use the middle mouse button.

Display

The main area of the Sequence Viewer displays the sequences of all the selected MSFs which are recognized as proteins and the selected sequences. Sequences can be read into QUANTA using the **Read Sequence/Alignment File** option on the **Sequences** pull-right under the **Files** pulldown (as described in Chapter 2.). By default, the MSF sequences are displayed above the non-MSF sequences. The residues from MSFs are colored the same as the C α atom of that residue in the molecule window and the sequence residues are colored according to one of the sequence coloring schemes, by default according to an hydrophilicity classification.

Identifying residues

When you pick any residue, its ID is reported to the textport, and, if **Highlighting** is on (See section below on icon functions) and the residue is in an MSF, then the sequence residue will be highlighted by a yellow box and the C α atom on the molecule will be highlighted with a yellow star.

Picking the residue in the sequence can be used for selection in many other situations. For example, to focus on a particular area of the molecule, choose the **Set Origin** tool from the **Protein Utilities** palette and then pick a residue on the **Sequence Viewer**. The C α atom of the picked residue will be set in the middle of the molecule viewing area.

Selecting the viewing area

The viewing area can be adjusted using the red slider bars below and to the left of the main viewing area. To move the slider, hold down the middle mouse button with the pointer over the slider and then drag in the required direction. To change the scale of the viewing area, hold down the shift key and the MIDDLE mouse button with the pointer over the slider bar. The dragging up/down or left/right will expand or contract the slider bar and the scale of the main viewing area.

Sequence names

The names of the MSFs or sequences appear to the left of the main viewing area. Clicking a sequence name toggles sequence activity off and on. The names of inactive sequences are colored grey. If the sequence corresponds to an MSF, the MSF activity, as shown in the Molecule Management Table, will also be updated. The sequence activity can also be updated using the **Activity** button to the lower left (labeled **A**).

Residue IDs

The residue IDs for every tenth residue are shown beneath the main viewing area. By default, these are the IDs for the first sequence in the table. The name of the sequence whose IDs are displayed is to the left of the IDs. Picking this sequence name will bring up a dialog box to enable selection of an alternative sequence to be labeled. The residue interval of the labels and

whether to include segment names in the IDs can be changed from the **Options** button to the lower left (labeled **O**).

Display of Graphs

Several applications automatically draw graphs to the Sequence Viewer. The Protein Design Secondary Structure Prediction module has options to plot **Hydrophobicity**, **Conservation** and **Composition**. These analyses are applied only to the currently active sequences and multiple plots may appear overlaid (e.g., the hydrophobicity of all active sequences will be generated by the **Hydrophobicity** tool).

Legend

Each plot is a different color and the legend to the left of the graph gives the name of the sequence or the parameter plotted in the appropriate color. Picking the plot name in the legend will toggle off and on the display of the plot. The legend for plots not currently displayed are colored grey. For some types of plots, the legend also includes a **Difference** option. When this is picked, the difference of two currently displayed plots will be shown. If there are more than two currently displayed plots, the difference will be between the first two.

Alignment

The plots are drawn with the parameter corresponding to a particular residue or column in the sequence alignment above the appropriate residue or column. If the sequence alignment includes gaps then there will be gaps in the graph plots for that sequence. When the tools in the Align and Superpose module are used to change the sequence alignment, then the graph plots will be automatically updated to keep in register with the sequences. If a difference plot is displayed then it will be updated with the alignment.

The Sequence Viewer icons

* Highlighting

Toggles on/off the cross highlighting of the residues in the Sequence Viewer and molecule window. If the tool is off, then the icon is colored grey. When the cross highlighting is on then picking either a residue of an MSF sequence in the Sequence Viewer or any atom of a residue in the molecule window will highlight the Sequence Viewer residue with a yellow box and the C α atom of the molecule with a yellow star.

O Options

Picking this icon brings up a dialog box with options to control the appearance of the Sequence Viewer. The heights of sequence and graph viewing area can be set — the size of the sequence viewing area is proportional to the number of sequences currently displayed up to some maximum number of sequences above which the height of the viewing area is constant. The

1.. The Sequence Viewer

interval of the sequence residue ID labeling can be changed from the default of ten residues and the inclusion of the segment ID in the label can be toggled off. The match symbol, a vertical yellow bar, is used in the Align and Superpose module to denote the homologous regions of sequence. The thickness and color of this annotation can be changed. The highlight annotation is a pale yellow box around a residue in the sequence viewer which indicates a residue selected while the **Highlight** option is on. The thickness and color of this annotation can be changed.

The residue interval of the sequence labeling can be changed and the inclusion of segment IDs in the label can be toggled on or off.

G Toggle Graph Display

This icon only becomes available after a graph has been drawn in the Sequence Viewer. Picking this icon will toggle off or on the display of the graphs and will contract the size of the Sequence Viewer window when the graph is toggled off. The same graph will be restored when it is toggled on.

>< Focus

After picking this tool, you should pick a residue on the Sequence Viewer or a molecule and the focus of the Sequence Viewer will be changed to place that residue at the center.

<> Expand

Expand the Sequence Viewer display to show the full sequences.

A Activity Selection

The activity of any individual sequence can be toggled by picking the name of that sequence on the viewer. To change the activity of several sequences you may find it more efficient to use the selection dialog box brought up by this icon.

D Display Selection

This icon brings up a dialog box in which you can select which sequences are displayed on the viewer and the order in which they are displayed. Any currently undisplayed sequences are listed in the left hand box entitled **Hide** in the order that they were read into QUANTA. The right-hand box entitled **Display in Order** lists all currently displayed sequences in the order in which they are displayed.

If a sequence is selected from the **Hide** list, it will be moved to the bottom of the **Display** list. If a sequence is selected from the **Display** list, it is moved to its appropriate place in the **Hide** list. The **Hide All** and **Display All** buttons will move all sequences to the appropriate list. It is possible to insert a sequence into the display list at a specified position by clicking the **Insert Above** button and then clicking the sequence above which the insertions should take place. The symbol $\longrightarrow\longrightarrow\ldots?$ will appear in the **Display** list and any sequence picked from either the **Hide** or **Display** list will move to that position. The **Insert Above** tool is switched off by clicking the **Hide** button.

? Find It

The user is prompted to enter a short amino acid sequence and the first occurrence of this short sequence in an active sequence will be highlighted by a red box in the Sequence Viewer. The icon is changed to ... and when

picked again will find the next occurrence of the sequence unless there are no more occurrences, in which case it will revert to ? and remove display of the highlight box.

1.. The Sequence Viewer

2. Reading and Writing Sequence Data Files

Overview

This chapter describes tools for importing and exporting sequence data. This chapter describes:

- Reading and writing sequence data files
- Demo of user data

Reading and Writing Sequence Data Files

The following options are accessed from the **Sequence** item on the **File** pulldown menu. These tools import and export sequence-only data which does not have associated atomic data. To generate an MSF with a full atomic representation of a sequence, use the **Create MSF from Sequence** tool in the Protein Editor utility. There is also an option to import sequence-related data which has been generated external to QUANTA for display in the sequence viewer and to output the sequence viewer as a Postscript file for printing.

Read Sequence/Alignment File

This tool reads in individual sequences in FASTA, EMBL/Swissprot or GCG (Wisconsin) format. These are described in appendix [F](#). It will also read in sequences from the alignment files in QUANTA alignment, GCG Pileup, GCG Pairs, or Clustal format.

If the **Restore Alignment** option is selected, then only the sequence alignment, and not the sequences themselves, will be read from the file and used to reset the alignment sequences within the viewer. If the **For Active Sequences Only** option is also picked, the alignment will be restored only for sequences which are currently active within QUANTA.

Note that the usual file extension for a Pileup file is msf (multiple sequence file), but to try to avoid confusion with the QUANTA MSF file, the file

2.. Reading and Writing Sequence Data Files

librarian will use a default Pileup file extension of pup. You will need to change the names of your Pileup files to use this extension or enter the required extension in the file librarian.

Read Sequence Data File

This tool reads data created outside of QUANTA and displays it in the Sequence Viewer graphs or uses it for coloring sequences. The data must be in either binary or ascii QUANTA sequence data format. These formats and how to generate them are described in detail in appendix A of this manual.

The file can contain multiple sets of data. Each data set has a label and the name of the sequence which the data applies to. The data sets associated with a sequence can either be used to color the sequence or can be plotted as a graph which is kept in synchronization with the associated sequence alignment.

The sequence data normally maps one datum per residue of a sequence, however it is possible to have data sets which are not associated with a sequence which map one datum for each column in the Sequence Viewer, but such data sets can only be displayed as graphs.

If either the **Plot Graph** or **Color Sequence** options is checked, then the data selection tools are presented. Two points to note:

- Within the input sequence data file it is possible to indicate which data sets should be plotted by default so, as a user, you might be presented with the best default selection.
- To color sequences using the sequence data files it is also necessary to have a seq_user_color.dat file which defines the color mapping for the sequence data. This file is described in more detail in appendix A..

Demo of User Data

To see how **Sequence Data Import** works you might like to try it with some demonstration files. This demonstration will read data output from the PHD package, which is a secondary structure prediction server at EMBL, into QUANTA and display the data within the Sequence Viewer.

Demo files are in \$QNT_ROOT/user_group_files/sequence_data are for the sequence of a dihydrofolate reductase (dfr) whose structure is known from crystallography, though this server would normally be used to make predictions for sequences for which the structure is not known.

The output from PHD is in the file dfr.phd and includes:

- The result of a sequence database search for homologous sequences. These sequences are aligned and written in the GCG Pileup format as part of the dfr.phd file. This alignment has been edited out from dfr.phd into the file dfr.pup.
- Performs an analysis of secondary structure prediction which reports a three state (helix, extended or loop) propensity. That is for each residue there is a probability value in the range 0-10 of it adopting each of the three states. This data will be plotted on a graph within QUANTA.
- For each residue in the sequence gives a prediction of the most probable secondary structure type which may be: helix, extended, loop or none of these. Within QUANTA the sequence will be colored to show the prediction.
- For each residue in the sequence gives a prediction of the most likely structural environment: exposed or buried. These predictions can be displayed in QUANTA by coloring the sequence.

In order to read this data into QUANTA, it must first be converted to QUANTA sequence data format by a quick program described in appendix A.. The file dfr.sqdat which is generated contains data sets which are labeled HELIX, EXTED, LOOP (the secondary structure propensities), SECSTR (the secondary structure prediction) and ACCESS (the predicted solvent accessibility). All five sets of data are associated with the sequence predict_h274 which is the original input dfr sequence.

There is also a seq_user_color.dat file which defines a suitable color mapping for the SECSTR and HPDACC data which can be used to color the sequence according to its predicted secondary structure or accessibility.

To run the demo:

1. copy these files from \$QNT_ROOT/user_group_files/sequence_data to your working directory:
 - seq_user_color.dat
 - dfr.pup
 - dfr.sqdat
2. Use the **Read Sequence/Alignment File** option to read the **Pileup** format file dfr.pup.
3. **Read Sequence Data** option to read the ascii data file dfr.sqdat. Select both the **Plot Graph** and **Color Sequences** options.

2.. Reading and Writing Sequence Data Files

4. In the **Color Residues According to Data** dialog box, select the SECSTR data set to color the sequence according to its final predicted secondary structure.
5. Finally, to see the accessibility prediction, use **Read Sequence Data File**. The original file name should still be selected by default. Make sure the **Color Residues** button is active and when presented with the **Color Residues According to Data** dialog box, select **HPDACC** as the data set.

Reference

Thanks to Burkhard Rost^{*} at the EMBL for allowing us to use the PHD server output for this demonstration.

The PHD server is at:

<http://www.embl-heidelberg.de/predictprotein/predictprotein.html>

Write Sequence File

This tool writes each currently active sequence to a separate file in EMBL, FASTA, PIR or GCG format. By default the filenames are derived automatically from the sequence name and the default file extension for that format. If a file of that name already exists then you will be warned and given the option to overwrite it or give an alternative name.

Plot Sequence Viewer

This produces a file in QUANTA plot format or Idraw Postscript format for printing the Sequence Viewer. The latter format can be used directly for printing or can be read into Idraw for editing. There is a check box to choose a color plot (currently only implemented for the Postscript format) and the output color should closely match the current QUANTA color. Adjusting the QUANTA colors using the **Color** dials will therefore adjust the postscript colors.

^{*}Rost, Burkhard; Sander, Chris: "Prediction of protein structure at better than 70% accuracy" *J. Mol. Biol.*, **232**, 584-599 (1993)

Rost, Burkhard; Sander, Chris; Schneider, Reinhard: "PHD — an automatic mail server for protein secondary structure prediction" *CABIOS* **10**, 53-60 (1994).

Rost, Burkhard; Sander, Chris: "Combining evolutionary information and neural networks to predict protein secondary structure" *Proteins*, **19**, 55-72 (1994).

By default, all currently displayed sequences will be plotted and any currently displayed graphs. By default, the entire range of the sequence is drawn. However, if the active range is currently selected then only that range is drawn. Since only short sequences will fit across a page the display of the viewer must usually be wrapped round. If the plot will extend over more than one page then each page will be written to a separate file and the files given the names name_0n.ps or name_0n.qpt where name is the file-name that you entered and n is the page number. If there are existing files with these names you will be warned.

Remove Sequence

Select sequences to close from the dialog box. Note that this will only close sequences and not MSFs.

3. Protein Utilities

Overview

The **Protein Utilities** palette consists of the common tools and options that are used with all of the Protein applications and utilities. When Protein Design, Protein MODELER, Protein Health or Profile Analysis are activated from the QUANTA **Applications** menu, the **Protein Utilities** palette is displayed.

This chapter describes:

- Simple representations of proteins
- Tools and options

References

D. Eisenberg, R.M. Weiss, T.C. Terwilliger, *Nature* **299**, 371-374 (1982)

D. Eisenberg, R.M. Weiss, T.C. Terwilliger, *Faraday Symp Chem Soc.* **17**, 109-120 (1982)

protein
utilities

Simple Representations of Proteins

The **Smoothed C α Trace**, **Secondary Structure** and **Hydrophobic Moment** tools on the **Protein Utility** palette are simplified representation of proteins which should help you to visualize the overall protein structure.

Secondary Structure Vectors

The vectors which are displayed by the **Secondary Structure** tool and used in several of the Protein Design utilities show the direction of a secondary structure element. The vectors are derived from the positions of the C α atoms of all the residues in that secondary structure element. The vector direction is the principle moment of the C α atom coordinates and the vector is positioned so that it goes through the average position of the C α atoms. The ends of the vector are the projection of the terminal C α atoms onto the vector.

Hydrophobic Moments

It is usually observed that the side chains of hydrophobic residues are oriented towards the interior of a protein. The hydrophobic moment of a residue is a vector whose length is proportional to the hydrophobicity of the residue and whose direction is dependent on the side chain orientation. The hydrophobic moment will generally point to the interior of the protein. The hydrophobic moments of several side chains can be summed to give some indication of the preferred orientation of the region of protein.

The hydrophobic moment vector of a residue is defined as having its origin at the position of the C α atom. The vector length is proportional to the hydrophobicity of the residue (as taken from standard amino acid hydrophobicity scales). The direction of the hydrophobic vector is found by averaging the vectors from the C α atom to all the non-hydrogen atoms in the side chain. If the residue hydrophobicity is negative (i.e., the residue is hydrophilic) the vector will point in the opposite direction to the side chain. Some hydrophobicity scales have all positive values, to use one of these for plotting hydrophobic moments the scale is adjusted by subtracting the average hydrophobicity value from the value for each individual amino acid.

The hydrophobic moment of a secondary structure element is defined as the sum of the hydrophobic moments for all residues in the element and the vector origin is the center of the secondary structure vector (i.e., the line you see drawn by the **Secondary Structure** tool).

Tools and Options

The **Protein Utilities** palette contains tools that are used during various functions in Protein Design, Protein MODELER, Protein Health, and Profile Analysis.

Select Active Range

This tool displays the **Pick Range** palette, you should pick the first and last residue of the required range on either the molecule or the sequence viewer. The selected range remains active, and the tool highlighted, until you pick the tool again.

Clear ID

This tool removes all atom identification labels that are displayed after picking atoms in the viewing area.

Atom Information

This tool prints information about a selected atom in the textport, such as atom name, atom number, and residue number.

Set Origin

This tool places the next picked atom at the center of the viewing area. The atom becomes the center of rotation for subsequent operations.

Center

This tool calculates and changes the geometric center of displayed atoms and places the molecule in the center of the viewing area.

Reset View

This tool resets the display so that all active and visible structures are completely viewable.

Distance

This tool displays the distance between two atoms picked in the viewing area.

Bond Angle

This tool displays the angle between three atoms picked in the viewing area.

<i>Dihedral</i>	This tool displays the dihedral angle of four atoms picked in the viewing area.
<i>Show Monitors</i>	This tool displays the labels from the geometry tools.
<i>Delete Monitors</i>	This tool removes the labels resulting from the distance, bond angle, and dihedral tools.
<i>Legend</i>	This tool toggles the display of the color legend in the viewing area that is located in the lower right corner of the viewing area.
<i>Smoothed CA Trace</i>	This tool displays a smooth C α trace through averaged coordinates from which the general fold of a protein is easily discerned. The color of the trace is taken from the current color of the C α atoms.
<i>Secondary Structure</i>	This tool displays the general protein structure and vectors for only active molecules. Color 4 (yellow) is used for strands; color 8 (purple) is used for helices.
<i>Hydrophobic Moments</i>	This tool displays hydrophobic moment vectors for active molecules. Residue vectors for hydrophobic residues are color 14 (pink); hydrophilic residues are color 12 (pale blue); secondary structures are color 3 (red).
<i>Torsion Table</i>	This tool displays the Torsion Table which list the torsion angles of all the active structures.
<i>Options</i>	This option displays the Protein Utilities Options dialog box for changing variables for smooth C α trace, secondary structure vectors, and hydrophobic moment and scales options.
<i>Molecule Colors</i>	This dialog box offers some protein-specific coloring and display schemes. The color schemes apply one color to each residue and have the same color for that residue in the structure and in the Sequence Viewer. The coloring on the molecule structure can be applied to every atom in the residue or to just the carbon atoms with the other atoms having their usual element color. This option can be toggled by checking the Color non-carbon atoms by element color type button. If the current coloring or display was not set up via this utility then the Color File or Display File buttons are checked.

Color by Structure Properties

There are two different sets of coloring options for structures and sequences since most of the coloring schemes for structure are not applicable to sequences. Molecules can be colored by structural properties: the secondary structure, structural domain, solvent accessibility and residue environment and there is an option to color protein structures by the **Sequence Classification**. These structure property coloring schemes are explained in the chapters describing the analysis of the properties (the residue environ-

ment coloring scheme is explained in the Profile Analysis chapter). If the secondary structure coloring mode is selected for a protein whose secondary structure is not known, then it will be derived. However, the other properties take significantly longer to calculate and the appropriate utility should be used to calculate the property before using the coloring mode. If information is not available for a particular coloring scheme, the structure will be given a neutral color.

Color by Sequence Properties

The sequence coloring modes of **Hydrophobicity** and **Size** color a residue according to a classification of amino acids types. The classification scheme is stored in the file \$HYD_LIB/protein_param.dat under the keyword CLASS. Users can amend the classification or add new schemes by editing this file — see the Protein Parameter File Appendix for further information.

Color by Homology

The homology coloring scheme is useful to show on the molecule structures and sequence viewer the regions of high and low homology. This complements the Match Residue tool in Align and Superpose, which highlights the homologous residues on the Sequence Viewer with vertical yellow bands. The criteria for homology is the same as is currently set in Align and Superpose by the Match Residues tool. The color scheme and the minimal score required for a residue to be colored as homologous can be changed in the Align and Superpose utility by selecting **Homology Color** from the **Options** dialog box.

4. Protein Editor

Overview

The Protein Editor provides tools to modify the sequence of a protein by mutating, inserting, or deleting residues. There are also tools to enter a new sequence, generate an MSF from a sequence, or change the hydrogen representation of the molecule.

This chapter describes:

- Editing proteins
- Tools and options

For more information see:

Protein User's Reference

- Create Homology Model
- Model Backbone
- Model Side Chain

Editing Proteins

The Protein Design Edit Protein utility has tools for mutating, inserting, and deleting residues. The same functions can be applied to MSFs or sequences.

protein
editor

Ideal Residue Definitions

The definitions of the composition and structure of amino acids as used in mutation, insertion, and regularization is taken from the structure definition file \$HYD_LIB/protein_structure.gsd. It is possible for users to add new residue definitions to the file or change existing ones — see Appendix D..

Regularization

After a structure has been modified, the conformation initially generated may be energetically unfavorable. For example, inserting an extra residue into a good structure is certain to be energetically unfavorable. In order to

accommodate the change, neighboring residues to the one edited may need to be moved. A regularization tool is provided which will attempt to find an energetically reasonable conformation, but note that it usually starts from a very poor conformation and can only find the local minima. For changes which involve inserting and deleting residues the tools in the Model Backbone utility probably need to be used.

Residues

Mutating residues

Mutated residues are given geometrically sensible structures and, as far as possible, the side chain torsions are copied from the old side chain. A local conformation optima will be found if the Regularization tool is active.

Inserting residues

Inserted residues are generated in a linear conformation appended to the residue before or after the insertion. Each new residue is given a default ID based on the preceding existing residue. For example, a residue inserted after residue 10 will get an ID of 10.1. If several amino acids are inserted, the fractional number is incremented by one for each: 10.2, 10.3, and so on.

Deleting residues

Deleted residues are simply removed from the structure. No other changes are made automatically. The sequence can be renumbered by the **Renumber Residues** tool.

Modeling side chains

After inserting new residues it is advisable to model the side chains. The **Auto Model** tool on the **Model Side Chains** palette will do this (see Chapter 9) and, to simplify the procedure, the Protein Editor utility automatically writes a selection file, `edit_side_sel.rsd` which lists the inserted residues.

Editing Segments

Segments can be created or merged using the **Atom Property Editor** on the **Edit** pulldown. The residue table in this utility has a column containing the segment name. If a segment name is changed then the edited residue, and all residues before it in the same segment, will be assigned to a new segment with the new name. To merge two segments, double click the segment column header to bring up a dialog box which has tools for merging segments.

Hydrogen Addition

Hydrogen atoms are added to those atoms with atom types which indicate that they are extended atoms. Extended atoms types are used to denote atoms which should have hydrogen atoms attached. For hydrogen addition

to work correctly the atoms must be correctly typed. The **Apply Dictionary** utility on the **Edit** pull-down retypes atoms if necessary.

Hydrogen atom coordinates are taken from templates in the file \$HYD_LIB/hydtpl.dat. There are templates for all the extended atom types commonly found in proteins. To derive the hydrogen coordinates, the guide atoms in the template are superposed over the extended atom and two neighbors in the protein and the hydrogen coordinates copied from the template hydrogen atom coordinates. The atom type of the extended atom is changed to the appropriate non-extended form.

Tools and Options

The Protein Editor interface consists of a conventional palette of editing tools and an amino acid selection palette.

Amino Acid Selection

This palette consists of a list of the standard amino acid types (non-standard types are listed in a scrolling list accessed via the **Non-standard** tool). As you enter a sequence of amino acids, they are displayed in the Sequence Viewer colored red.

Non-standard

Presents a scrolling list of all amino acids in the protein structure definition file which are not one of the twenty standard amino acids.

Keyboard

This tool displays a dialog box to enter the sequence as a string of one letter amino acid codes from the keyboard.

Undo Last

This tool removes the last amino acid entered. The tool can be used repeatedly.

Quit

This tool exits from the palette without saving any sequence entered, so no insert or mutate action is applied.

Finish

This tool exits from the palette and applies the insert or mutate action.

The Editing Tools

Regularize

If this tool is active, then after each mutate, insert or delete function the structure will be regularized. The regularization tool is described more fully in chapter 8. By default, only the changed residue is regularized after mutation but after insertion or deletion two residues on either side of those changed will also be regularized. The number of residues regularized can be changed by the **Regularization Options** tool.

4.. Protein Editor

Regularization Options

The regularization method and options are described in the *Model Backbone* chapter.

Use No Hydrogens/Use Polar Hydrogens/Use All Hydrogens

When one of these tools is selected then all active structures will be converted to the designated hydrogen representation.

Insert Before/Insert After

These tools insert residues before or after a residue. Once picked, these tools remain active and highlighted until they are toggled off or another editing tool is toggled on. After choosing this option, you pick a residue on the molecule or **Sequence Viewer**, and then choose one or more residues to insert before or after the specified residue.

Mutate

While this tool is active you can pick a residue on the molecule or **Sequence Viewer** and then select a new residue type from the amino acid selection palette. Once picked, this tool remains active and highlighted until it is toggled off or another editing tool is toggled on.

Mutate Range

This tool mutates a specified range of the sequence. The **Pick Range** palette enables you to select the range of residues to mutate. The **Amino Acid Selection** palette is displayed and the changed residues are shown in red in the **Sequence Viewer**.

Delete

While this tool is on you may pick a residue on the molecule or **Sequence Viewer** and it will be deleted. Once picked, this tool remains active and highlighted until it is toggled off or another editing tool is toggled on.

Delete Range

This tool activates the **Pick Range** palette from which a range of residues can be selected for deletion.

Change Terminal

Select one terminal residue on either the molecule or the Sequence Viewer and a palette like the amino acid selection palette (but with the appropriate terminal groups) will be displayed, allowing you to select a new group. It is not appropriate to change the terminal of a sequence.

Disulfide

You should select a cystine residue by picking from either the Sequence Viewer or by picking one atom in the residue on the molecule. If that cystine is already part of a disulfide bond, then that bond will be broken. If it is not disulfide bonded, then it will make a bond to any neighboring cystine residue. It is not appropriate to make or break disulfides in sequence only data.

Renumber Residues

All residues in a range are numbered consecutively from the first residue ID. This tool activates the **Pick Range** palette to pick a range of residues. You are then prompted for the ID of the first residue in the range, which, by default, is the current ID. If an insertion code is entered, all residues in the range are given the same ID as the first residue and they are given incremental insertion codes which follow from the entered insertion code.

Create Sequence

The amino acid selection palette is presented and you can enter a sequence which will be displayed in the Sequence Viewer below any existing sequences. On exiting the amino acid selection palette you are prompted for a name for the new sequence.

Create MSF from Sequence

By default, MSFs will be created for all currently active sequences and by default the new file is given the same name as the sequence. The sequence is removed from the selection and replaced by the MSF.

Finish

This tool exits from the palette. If any structures have been edited but not saved to MSF you will be prompted to save them. If they are not saved then the structures will revert to those currently saved in the MSF.

4. Predict Secondary Structure

Overview

This utility is concerned with sequence analysis. It can be used for sequences for which the structure coordinates are not defined. There are six different types of analysis that can be displayed, three of which are prediction methods. The results of the analyses are usually presented as plots of property versus sequence position above the sequences in the Sequence Viewer. When there are gaps in the sequence alignment, then there will normally be gaps in the plots.

This chapter describes:

- Predicting secondary structures
- Tools and options

References

- L. H. Holley and M. Karplus, *Proc. Natl. Acad. Sci., USA*, **86**, 152-156 and G. L. LaRosa et al, *Science* **249**, 932-935.
- J. Garnier, D. J. Osguthorpe and B. Robson, *J. Mol. Biol.* **120**, 97-120 (1978).
- G.D. Rose et. al. *Science* **229**, 834-838 (1985).
- J.L. Fauchère and V. Pliska *Eur. J. Med. Chem - Chim.* **18**, 369-375 (1930).
- D. Eisenberg et. al. *Faraday Symp. Chem. Soc.* **17**, 109-120 (1982).

Predicting Secondary Structures

Several methods are available to predict the secondary structure of a sequence. The three predictions that are used in Protein Design are the Momany, GOR, and Holley/Karplus methods of prediction. In addition to these methods, this module also provides tools to plot the hydrophobicity profile and conservation profiles on the active molecules.

Secondary structure prediction methods usually consider three classes of secondary structure: α -helix, β -stand and 'neither of these'. Some methods may have a turn classification. Most methods derive, for each residue in the sequence, a probability, or propensity, of the residue occurring in each of the secondary structure types. The calculated propensities are plotted in the Sequence Viewer. The predicted secondary structure type for each residue

4.. Predict Secondary Structure

is the type with the highest propensity with some allowance made for the fact that secondary structure elements are of some minimal length.

Momany Prediction

This prediction modifies the Zimm/Bragg method. The Zimm/Bragg method, which is based on the classical Chou-Fasman technique, was developed by Dr. Harold Scheraga and co-workers Momany, Lewis, and Zimmerman.*

The Zimm/Bragg method has two coefficients, a one for helices and a zero for non-helices. Momany modified this method by enhancing these parameters with additional values, so when specific characteristics were found in sequence, such as turns and anti and parallel β -sheet regions, the value of the coefficient increased.

This method makes an initial pass through the primary sequence to determine the Zimm/Bragg coefficients. Subsequent passes are then made to enhance the coefficients by identifying certain patterns found in the primary sequence.†

For example, in the initial pass on a sequence there are regions found and categorized as being helical. On the subsequent pass it is noted that several of those regions have polar residues separated by two residues, indicating a classical 1-4 helical arrangement. Therefore, the coefficients for those polar residues would be enhanced for their 1-4 helical character.

After multiple passes, the resulting prediction coefficients are normalized and used for the final prediction of helical, β -strand, and turns.

Holley/Karplus Prediction‡

This prediction is based on a neural network that identifies three secondary types: helix, strand, and coil. This neural network is trained on 48 unrelated proteins. Its method for assigning a secondary structure uses a window of 17 residues to determine the central residue, recognizing that a residue may be affected by another residue eight places away in the sequence. This is implemented within QUANTA as a translated neural net.

Once the assignments have been made they are smoothed such that:

*This was developed in the late 60's and early 70's. Momany, Lewis, and Scheraga developed work for bends and Momany, Scheraga, and Zimmerman developed work for helix work.

†This is based on work of Fred Cohen et al. at UCSF.

‡L. H. Holley and M. Karplus, *Proc. Natl. Acad. Sci., USA* **86**, 152- 156 and G. L. LaRosa et al, *Science* **249**, 932-935.

- Sheet regions are a minimum of two residues.
- Helix regions are a minimum of four residues long.
- Shorter regions revert to coil.

GOR Prediction*

This method identifies four secondary structure types: helix, extended, reverse turn, and coil. It uses an analysis of a 17-residue window to determine the secondary structure of the central residue; the residues at the center of the window have greatest influence.

The parameters used in this method are derived from a statistical analysis of protein structures to determine the probability of each amino acid type occurring at each position in a 17 residue window around a residue of each of the four secondary structure types.

The prediction can be weighted for a particular type by varying the decision constant. This constant is subtracted from the score for a weighted secondary structure type.

Conservation Profiles

The calculation of conservation profiles uses the table identified with the label CONSERV in the file \$HYD_LIB/protein_seq_param.dat. This table defines 10 classes of amino acid (e.g. small, aromatic, acidic) and specifies which amino acids belong in which class. The degree of conservation between two amino acid types is the number of classes to which they both belong divided by 10 so the maximum conservation score is one. The conservation value of a column of aligned residues in the sequence table is the sum of all the pairwise conservation comparisons divided by the number of comparisons. The maximum value is one. The conservation profile can be smoothed by averaging over a range of residues. The window length is set by the **Profile Options** tool.

Hydrophobicity Scales

Hydrophobicity is a measure, for each of the amino acids, of its immiscibility with water. Generally apolar amino acids have higher hydrophobicity parameters and are more likely to occur on the interior of proteins rather than exposed to solvent.

*J. Garnier, D. J. Osguthorpe and B. Robson, *J. Mol. Biol.* **120**, 97-120 (1978).

4.. Predict Secondary Structure

There are three hydrophobicity scales used: the Rose^{*}; Fauchère and V. Pliska[†]; and Eisenberg[‡] scales. The parameters are stored in the file \$HYD_LIB/protein_seq_param.dat and alternative parameter sets can be added to that file.

The Rose scale is based on the statistical analysis of the environment of protein crystal structures. The Fauchère and V. Pliska scale determines the free energy of transfer of amino acid analogs between octanol and water. The Eisenberg consensus scale is an average of several other scales.

Hydrophobicity is usually analyzed by averaging over a fairly long window (e.g. in range 7 to 21 residues) and regions of low hydrophobicity are generally found to be loop regions of the protein which are exposed to solvent.

Sequence Viewer Plots

All of the parameters analyzed in this utility are plotted in the sequence viewer above the sequences. The parameters are usually calculated for all active sequences and the plot is aligned to the sequence so there may be gaps in the plot where there are gaps in the sequence alignment. If you exit the Prediction **Utility**, enter the Align and Superpose utility, and uses any of the tools there to change the sequence alignment then, where appropriate, the plot will be updated to keep in sync with the sequence. The hydrophobicity plot might be a useful in alignment as, generally, the hydrophobicity plots of two homologous structures are strongly correlated.

The plot legends are colored the same as the plot that they identify. By picking a plot legend you can toggle on or off the display of the plot. The legend for an undisplayed plot is colored gray. To change the display status of several plots it may be quicker to pick the plot title (at the top of the legend) and a selection dialog box is presented. To toggle on or off the display of all plots pick the **G** icon on the bottom left of the sequence viewer.

The secondary structure prediction tools are applied to all active sequences and the sequences recolored according to their predicted secondary structure. The secondary structure propensities for one sequence will be plotted in the Sequence Viewer. If there is more than one sequence active, then you are prompted to select one sequence for which propensities are plotted.

^{*}G.D. Rose et. al. *Science* **229**, 834-8388 (1985).

[†]J.L. Fauchère and V. Pliska, *Eur. J. Med. Chem. -Chim. Ther.* **18**, 369-375 (1930).

[‡]D. Eisenberg et. al. *Faraday Symp. Chem. Soc.* **17**, 109-120 (1982).

Saving Predictions

Predictions are automatically saved to a file which is given a name of the form *sequence_method_predict.out* where *sequence* is the sequence name and *method* is the prediction method. For MSFs there is an option to save the predicted secondary structure to an MSF as extra information.

Tools

Plot Hydrophobic Profile...

This tool plots the hydrophobic profile for the active molecules.

Profile Options...

This opens the Hydrophobic Profile Options dialog box

You can change the hydrophobicity scale and the window length used in the hydrophobicity plot. The window length for the conservation plot is also changeable.

The options in this dialog allow you to select different scales, residue window lengths, molecules, drawing averages, and difference profiles

Plot Conservation Profile

This tool plots the conservation profile for two or more active molecules. The conservation profile is a measure of the extent of sequence conservation along two or more aligned sequences. Sequences must first be aligned. A high conservation number is given when similar chemical types of amino acid occur at a position, and a lower number is given when chemical types differ. Conservation profiles can be averaged over a window - the length can be changed via the **Profile Options** tool.

Plot Composition

This plots a graph in the sequence viewer showing, for each column of residues in the sequence viewer, the number of residues which fit into a given chemical classification such as acidic or aromatic. The classes are defined in \$HYD_LIB/protein_seq_param.dat under the keyword CONSERV. Inactive sequences are excluded from this analysis. The plot shows ten different classifications and may be difficult to interpret when all classifications are displayed simultaneously. You can toggle off or on the display of a given classification by picking its name on the plot legend. To change the display status of multiple classifications pick the legend title **Composition** and you will be presented with a dialog box.

Plot Momany Prediction

This tool performs a Momany secondary structure prediction for each active sequences and recolors the sequences according to the predicted secondary structure. Each prediction is written to a file of the form *sequence_momany_predict.out*. The prediction propensities for one sequences are plotted in the sequence viewer.

4.. Predict Secondary Structure

<i>Plot GOR Prediction</i>	This tool performs a GOR prediction based on all the currently active sequences. To get meaningful results the sequences must be aligned. The prediction is written to a file of the form <i>sequence_GOR_predict.out</i> . The prediction propensities are plotted to the sequence viewer.
<i>GOR Options</i>	This tool opens a dialog box that allows you to reset the ranges and variables for the GOR prediction.
<i>Plot Holley/Karplus Prediction</i>	This tool performs a Holley/Karplus Prediction for each active sequence and recolors the sequence according to predicted secondary structure. The predictions is written to a file of the form <i>sequence_holley-karplus_predict.out</i> and the prediction propensities plotted to the sequence viewer.
<i>Edit Secondary Structure</i>	This tool allows you to change the secondary structure assignment for a single residue or range of residues. The mode of residue selection is determined by the Pick Residue and Pick Residue Range tool. Once the residue or residue range has been chosen, the Secondary Structure dialog box opens. Secondary structures can then be reassigned to the specified areas.
<i>Pick Residue</i>	This allows you to select single residues in order to edit their secondary structures.
<i>Pick Residue Range</i>	This allows you to select a range of residues in order to edit their secondary structures.
<i>Save to MSF</i>	Predictions for sequences which are from MSF files can be saved as secondary structure extra information in the MSF. You will be prompted to give the data a label. Be careful not to confuse predicted secondary structure with that derived from analysis of the structure.
<i>Read from MSF</i>	Saved secondary structure predictions can be restored from the MSF file.
<i>Finish</i>	This tool exits the palette. You will be prompted to save any unsaved secondary structure predictions to the MSFs.

5. Align and Superpose

Overview

The **Protein Design Alignment** palette provides options and tools for aligning, matching, and superposing proteins. Sequence alignment can be done automatically or the alignment can be edited manually. There are tools to aid alignment: dot plots, alignment constraints and graphical indications of homology. The alignment and matching of homologous sequences can be based on a variety of sequence and structural criteria.

This chapter describes:

- Aligning and superposing
- Superposing structures
- Tools and options

References

S. B. Needleman and C. D. Wunch, "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins," *Journal of Molecular Biology*, **48**, 443 (1970).

M. O. Dayhoff, Atlas of Protein Sequence and Structure (National Biomedical Research Foundation, Silver Spring, Md., 1978), **5**, supplement 3.

D. F. Feng, R.F. Doolittle, "Progressive Sequence Alignment as a Prerequisite to Correct Phylogenetic Trees," *Journal of Molecular Evolution*, **25**, 351-360 (1987).

Aligning and Superposing Sequences

A variety of options are available for aligning sequences, matching residues, and superposing structures. A general discussion describing the different algorithms and options used for these tools follows.

Using Active Sequences and Active Ranges

All the tools on this palette are only applied to active sequences. So to align, match or superpose a limited set of the sequences or molecules you should change the sequence activity. Sequence activity is indicated on the Sequence Viewer by graying out the names of inactive sequences. The

align and
superpose

5.. Align and Superpose

activity can be changed by either picking the sequence name on the sequence viewer or picking the **A** icon on the bottom left of the Sequence Viewer to bring up a dialog box which will allow you to change the activity of multiple sequences more efficiently. If the sequence is also an MSF then its activity can be changed in the Molecule Management Table.

Automatic alignment, manual editing of the alignment and application of **Undo All** and **Undo Last** can be applied only to residues within the active range. This can be particularly useful when manually editing small regions of the alignment: the **Active Range** tool can be used to ensure that the rest of the alignment is unchanged by insertion and deletion of gaps in the active range. In order to maintain alignment of residues to the right of the active range, gaps may be inserted or deleted at the right hand end of the active range.

The match tools are also applied only to residues within the active range. The active range can be set by picking the **Set Active Range** tool on the **Protein Utilities** palette and is indicated on the Sequence Viewer by triple red lines showing its limits. The thicker, innermost line is the actual limit.

Criteria for Aligning and Matching Sequences

Sequence alignment algorithms attempt to align pairs of residues which are similar. The algorithms require some quantitative measure of similarity. Conventional sequence alignment uses an amino acid substitution matrix which has been derived from analysis of amino acid substitutions observed in families of proteins through the course of evolution. It is possible to use other criteria; particularly if the structure of the protein is known, there are criteria for aligning the residues in the equivalent positions and environments in the structure.

The same criteria which are used to optimize the alignment can also be used to indicate the degree of homology between proteins. The match tools in this utility identify the homologous residues or ranges of residues.

In this utility there are five possible scoring schemes for alignment and matching and you can use weighted combinations of these schemes. The default scoring is a conventional sequence homology scoring system.* It is also possible to use some combination of these criteria — a combination of 50% sequence similarity, 30% secondary structure similarity and 20% Cα-Cα distance criteria is useful for recognizing homologous structures.

Sequence homology

The sequence homology scoring system uses the conventional Dayhoff amino acid substitution matrix which is based on the probability of replac-

*M. O. Dayhoff, *Atlas of Protein Sequence and Structure* (National Biomedical Research Foundation, Silver Spring, Md., 1978), 5, supplement 3.

ing one amino acid type by another as observed in the evolution of families of proteins.*

Secondary structure homology

The secondary structure homology scoring system scores favorably for aligning residues of similar secondary structure and penalizes aligning non-similar secondary structure.[†]

Residue accessibility homology

The accessibility scoring uses the residue fractional solvent accessibility. The score is linearly dependent on the difference in the fractional accessibility. There is a maximum score of 10.0 for no difference in accessibility. The score decreases linearly to zero for a cutoff difference of 0.3. The maximum score and cutoff can be changed using the **Alignment Scores** tool.

Residue Environment

The environment class of a residue is that defined by the method of Luthy, Bowie and Eisenberg[‡] as used in Profile Analysis and is based on the solvent accessibility, polarity of environment and secondary structure of the residue. Before using the environment class as a criteria for alignment it should be calculated for all the relevant structures using the **Plot Structure Profile** tool in the Profile Analysis application^{***}.

C α -C α distance homology

The C α -C α distance homology scoring system is based on the interatomic distances between the C α atoms of aligned residues in different sequences. This scoring system is only applicable after structures are superposed. The score for aligning a pair of residues is linearly dependent on the C α -C α distance, there is a maximum score of 10.0 for a distance of zero. The score decreases linearly to zero for a cutoff distance of 5.0Å. The maximum score and cutoff can be changed using the **Alignment Scores** tool.

Alignment

The conventional pair-wise sequence alignment method described by Needleman and Wunch^{††} aligns two sequences to maximize the alignment score. The alignment score is the sum of the scores for all pairs of aligned residues, minus an optional penalty for the introduction of gaps (automatic insertions and deletions) into the alignment. If there are more than two sequences to be aligned then they are aligned chronologically in pairwise fashion.³

*These scores are stored in the file \$HYD_LIB/protein_align_score.dat.

†These scores are stored in the file \$HYD_LIB/protein_align_score.dat.

‡R. Luthy, J.U. Bowie & D. Eisenberg "Assessment of protein models with 3D profiles" *Nature* **356**, 83-85 (1992)

***The scoring schemes stored in the file \$HYD_LIB/protein_align_score.dat.

††S. B. Needleman and C. D. Wunch, "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins", *Journal of Molecular Biology*, **48**, 443 (1970)

5.. Align and Superpose

To align more than two sequences an alignment is performed for all pairwise combinations of active sequences and the alignment score which indicates the degree of homology of the two sequences is calculated. The normalized alignment score is also calculated by multiplying the score by 100 and divided by the number of residues in the shorter of the two sequences.

The normalized alignment score for each pair of sequences is reported in the textport and plotted as a dendogram.* This dendogram indicates the relationships and order in which pairwise alignment is used to align multiple sequences. Sequences that join at the leftmost node in the dendogram correspond to the highest normalized alignment score, and therefore are the most similar sequences so they are aligned first. After a pair of sequences are aligned with each other they are kept fixed with respect to each other and aligned against more dissimilar sequences.

As the iterative alignment procedure is performed the alignment scores are reported in the textport. The Sequence Viewer is updated, showing the new alignment.

Gap Penalties

Gap penalties weight against the alignment algorithm introducing insertions and deletions into the alignment. Alignment can be significantly affected by the size of the gap penalties used, particularly in cases of low homology.

There are three forms of gap penalty used in QUANTA

- penalty for initially opening a gap
- penalty for every extra residue added to the gap
- penalty for mismatching the ends of the sequence

The effect of the first two forms is that a large penalty weighs against opening a gap and there is a smaller additional penalty applied for extending the gap. The penalty for mismatching ends is, by default, lower.

The default opening penalty is a fixed value but there is an alternative penalty scheme which is dependent on secondary structure. Opening a gap in the middle of a secondary structure element, helix or strand, is heavily penalized which openings at the end of the element are less heavily penal-

*A *dendogram* is a plot showing the family tree of three or more sequences and is based on scores from pairwise comparisons of sequences done either by the Align Sequences or the Match Residues tools. A dendogram plot will be produced automatically if three or more sequences are aligned or if the **Dendogram** option is checked in the Match Residues tool. A dendogram is like a family tree diagram showing the family relationship between sequences with most similar sequences connected by the shortest branches.

ized. This tends to force insertions and deletions to loop regions which is where they are observed most commonly in practice.

If an initial alignment does not produce the expected results then it may be worthwhile to experiment with the gap penalties whose values can be changed using the **Options** tool.

Manual Alignment Editing

There are several tools to allow manual adjustment of the alignment. These are particularly useful when used in conjunction with the **Match Residues** option to **Update match when alignment changed** so that as the alignment is changed the match bars in the sequence table and/or the match score plotted in the Sequence Viewer give feedback on the quality of the alignment.

The two basic manual alignment editing methods are addition and removal of gaps which allow shifting of sequences by one residue position at a time. To make bigger shifts to a sequence the **Align Two Residues** tool can be used. Moving entire sequences can be done with the click and drag facility in the Sequence Viewer.

By default, adding and removing gaps will cause a readjustment of the alignment for all the positions to the right of the edited position. The changes can be limited by setting an active range using the **Set Active Range** tool on the **Protein Utilities** palette. Changes to the alignment will not be propagated outside of the active range. Gaps may be inserted or deleted at the right hand end of the active range in order to maintain alignment of residues to the right of the active range.

Saving and Restoring Alignments

The **Undo Last** and **Undo All** tools will allow backtracking after automatic or manual alignment. Alignments can also be saved to file and restored later by using the Sequence Data utility on the **Files** pulldown. Use the **Write Alignment File** option to save the alignment and the **Read Alignment/Sequence File** option to restore the alignment. The **Restore Alignment Only** option should be selected so the sequences are not read into QUANTA again. Any file format can be used but the Clustal format is similar to that used by QUANTA to save alignments between sessions. Note that if you do not want to restore the alignment for all of the sequences then some sequences can be made inactive and the **For Active Sequences Only** option should be checked.

Dot Plots

Dot plots show a comparison between two sequences and can provide useful feedback on the quality of an alignment and suggest alternative alignments which might be tested. The x axis of a dot plot is the residue position of the first sequence and the y axis is the residue position of the second sequence. The value shown at the position x,y in the plot is a comparison score for residue x of sequence 1 to residue y of sequence 2. Various parameters can be scored and plotted: the default is the amino acid similarity score as given by the Dayhoff comparison matrix. It is convention to show normalized rather than absolute scores on dot plots. To do this, the mean and standard deviation of the scores for all the points on the dot plot are calculated and then only the relatively high scores are shown in terms of their number of standard deviations above the average score.

Path of alignment

Also shown on the dot plot is the path of the current alignment. This is the blue line with small blue dots showing where residue x is aligned to residue y. For two similar sequences the alignment path will run roughly diagonally across the plot from bottom left to top right. Where there are gaps in the alignment, the alignment path will not run parallel to this leading diagonal and there will be a relatively long gap between the small dots indicating aligned residues.

Dot plots are usually drawn to show the comparison of a range of residues rather than single residues. If you are unfamiliar with dot plots then try drawing a dot plot for two short similar sequences. Before doing so, however, set the dot plot window to one and switch off the normalization of dot plot scores (use the **Options** tool to access the Dot Plot Options dialog box). This shows, for the purposes of comparison, the score of every individual residue in sequence one against every individual residue in sequence two —the checker board effect is very difficult to interpret.

Attempting dot plots of various window lengths

Now try dot plots with window lengths of three, five and eleven. Using longer window lengths, you will see diagonal lines appear on the plot and a strong band along the leading diagonal if the two sequences are significantly similar. If the two sequences are aligned automatically, then the alignment path shown on the plot would be expected to overlay the strong diagonal lines.

In calculating a dot plot with window length eleven, the sum of the scores for comparison of eleven consecutive residues in both sequences one and two is assigned to the position on the plot corresponding to the sixth residue in the comparison window of each sequence. The average and standard deviation of the scores for all points on the plot are then calculated. Where the comparison score is above the cutoff a dot is drawn on the dot plot for each pair of residues in the comparison windows. This gives the diagonal lines which you see on the dot plot. Since each residue contributes to mul-

multiple comparison windows it is possible that it will contribute to more than one comparison scores which is above the display cutoff; when this happens the position on the dot plot is colored to indicate the larger of the comparison scores. Overlap of comparison windows with good scores may also give diagonal lines on the plot which are longer than the window length.

Dot plots for similar sequences show a strong diagonal trace roughly along the leading diagonal and after automatic alignment the blue line showing the alignment path will follow this trace. If there are other strong traces close to the leading diagonal then they indicate possible alternative alignment paths which can be explored using constraints.

Dot plots can be drawn for a limited region of two sequences by using the **Select Active Range** tool on the Protein Utilities palette. This can be useful in analyzing a region of low homology. Using a shorter window length will also be useful in this situation.

Alignment Constraints

Alignment constraints enable you to bias the automatic alignment algorithm to align the constrained residues. Constraints might be needed if there is experimental evidence for alignment of certain residues or if you want to explore non-optimal alignments as suggested by the dot plot. When constraints are used in an alignment, a large favorable score is assigned to aligning the constrained residues. This does not absolutely guarantee that the algorithm will align the constrained residues — the penalties incurred by aligning inappropriate residues and the gap penalties may outweigh the constraint weighting. To enforce the constraints, it is possible to increase the constraint weighting, but it is probably better to also assign constraints to neighboring residues.

Matching Residues

Matched residues are aligned residues (i.e., residues in the same column of the sequence viewer) which are homologous. The degree of homology may be determined by a variety of criteria:

- The amino acid type is identical or similar.
- The residue environments, accessibility or secondary structure are similar
- The residues are close in space (this is the $C\alpha$ - $C\alpha$ distance criterion).

Matched residues

Matched residues are usually indicated on the sequence viewer by a vertical yellow line. The appearance of this line can be controlled in the **Sequence Viewer Options** tool (accessed through the **O** icon at the bottom left of the

5.. Align and Superpose

Sequence Viewer). An alternative means of display is to plot the match score as a graph in the Sequence Viewer. An option to update the matched residues whenever the sequence alignment is changed is on by default.

The match scores can be analyzed to give pairwise comparison scores between all the active sequences and these scores can be analyzed to generate a dendrogram of the family relationship between all of the sequences based on whatever criteria is currently being used for the match analysis.

You may also select the matches manually, an option which is useful when the matched residues are to be used as the selection criteria in another function such as in the **Copy Matched Residues** tool in the Create Homology Model tool.

The match score is calculated on a column-by-column basis, using the current alignment of the active sequences. Alternatively, the score can be averaged over several columns around the column under consideration. This averaging of match scores is useful for identifying homologous regions rather than just similar individual residues. The match window length is controlled by the **Match Options** tool.

RMS Deviation tool

Another tool which provides information to help assess alignments is the **RMS Deviation** tool. There is an option to plot a graph in the sequence viewer of the distance between equivalent atoms in aligned residues.

Color by Homology

The homology between sequences is indicated on the Sequence Viewer by vertical yellow bands and can be shown on two molecule structures by dashed lines between matched residues but this latter presentation is not easily interpretable for more than two structures. Coloring the structure residues according to their homology is useful in this case and can be activated by the **Color by Homology** option in the **Molecule Color** on the **Protein Utility** palette. The coloring ranges used by this tool can be changed using the **Color by Homology** option accessed via the **Options** tool.

Superposing Structures

Structure superposition overlays atoms within the matched residues of the active structures, using a least squares algorithm. By default only the C α atoms are superposed but alternative selections are available using the **Superposition Options** under the **Options** tool.

To superpose multiple molecules, there are several cycles of superposition.* In the initialization cycle, each of the other molecules are superposed

onto a target molecule (by default this is the first selected molecule). For subsequent cycles, a template, which is an average of all molecules, is calculated, and each molecule is superposed onto the template. For each cycle, the root mean square (rms) difference in atomic coordinates between each molecule and the target template is reported. After each cycle, a new average template is calculated, and the rms difference in coordinates between this template and the template from the previous cycle is reported. If only two molecules are being superimposed, the rms difference reported is one half the rms difference between the two molecules.

If the RMS difference in template coordinates between cycles is less than 0.1Å, then the refinement is terminated; otherwise, it is terminated after 10 cycles. If you have opted to output the transformation matrix (see under the **Options** tool), then the translation vector and rotation matrix that have been applied to the coordinates of the molecule in order to bring it to the final superposed position are reported.

After the structures are superposed the interatomic distances between the C α can be used as a criteria in alignment and this can be a useful means of refining the alignment to reflect structural homology.

Tools and Options

Align Sequences...

This tool aligns all the currently active sequences. If the **Select Active Range** tool has been used then the alignment will only be applied to the active range. The default alignment criteria is to align similar residues types but the **Alignment Weights** tool can be used to change the criteria. When there are only two active molecules, the sequences are immediately aligned.

To align more than two sequences the usual protocol is to align all possible pairwise combinations of sequences and calculate an alignment score. Cluster analysis of these scores determines the family relationship between the sequences which is represented by a dendogram. The default protocol then aligns all the sequences in an order determined from the dendogram. Alternative to this default protocol you may stop after generating the dendogram or you may select two sets of sequences to align. A dialog box presents you with these options when you align more than two sequences:

The options for alignment are:

- Do automatic alignment of active molecule. This does the default alignment.

*This follows the method of Sutcliffe et al (M.J. Sutcliffe, I. Haneef, D. Carney and T.L. Blundell, *Protein Engineering* **1**, 377-384 [1987]).

5.. Align and Superpose

- Do pair-wise alignment and cluster analysis. This calculates the pair-wise alignment scores, does a cluster analysis and generates a dendrogram. The actual sequence alignment is not performed.
- User selection of molecule sets to align. A dialog box is displayed for you to select two sets of sequences. An alignment is performed which keeps all of the sequences in one set in the same alignment to each other and aligns them to all of the sequences of the other set.

Alignment Weights...

This tool displays the Alignment Weights dialog box which allows you to choose a weighting scheme for using a combination of the different homology criteria. All weights should be in the range 0.0 to 1.0

Alignment Scores...

This option displays the Score Parameters dialog. This dialog allows you to specify score parameters, cutoffs and change the align score file.

Weight to fix constrained residues in alignment

(default 100) The **Constraint** tool is used to select residues which will be pulled into alignment by the automatic alignment. The weighting of the constraint can be changed through this option.

Maximum distance score

(default 10.0) and

Distance cutoff

(default 5.0). These parameters affect the C α -C α distance homology scoring. The maximum score is given for a distance of zero and the score decreases linearly to zero for the cutoff distance.

Maximum accessibility score

(default 10.0) and

Cutoff for difference in residue accessibility

(default 5.0) These parameters affect the accessibility homology scoring. The maximum score is given for an accessibility difference of zero and the score decreases linearly to zero for the cutoff difference in residue accessibility.

Protein Align Score File

By default the residue type scoring scheme is taken from the file \$HYD_LIB/protein_align_score.dat which contains the Dayhoff substitution scoring matrix. An alternative file name can be entered here. Note that the file should have the same format as the default file.

Undo Last

This tool undoes the last sequence alignment or alignment edit.

Undo All

This tool remove all gaps from the active sequences. This only applies within the **Active Range** if it is on.

Add Gap

When this tool is active, you can pick residues (on the Sequence Viewer or active molecules) and add a gap before that residue. The tool remains highlighted and active until it is deselected.

Delete Gap

When this tool is active, you can pick a gap on the Sequence Viewer to delete it. The tool remains highlighted and active until it is deselected.

<i>Align Two Residues...</i>	This tool aligns two residues from the sequences of two different active molecules. You are prompted with the Pick Residue palette and should then select two residues on the Sequence Viewer or molecules. The left-most of the two residues will be moved into line with the rightmost.
<i>Dot Plot...</i>	This tool calculates and displays a dot plot for two sequences. If more than two sequences are currently selected then you are prompted to select just two. If the Active Range is on, then the dot plot is drawn for only the active range.
<i>Match Residues</i>	This tool brings up a dialog box with the option to choose the match criteria and also with options to control the mode of action. These are:
<i>Undo All Matches:</i>	Undo all display of matches.
<i>Select matches:</i>	a residue selection palette allows you to manually select matched residues.
<i>Update matches when alignment changed:</i>	While this option is on the Match Residues tool on the palette will remain highlighted and the matched residues will be recalculated every time the alignment is changed.
<i>Plot graph of match scores:</i>	The match scores are plotted in the Sequence Viewer. This option can be used in conjunction with the previous one to give updates of the plot as the alignment is changed.
<i>Plot dendogram:</i>	Plot a dendogram based on the pairwise inter-sequence match scores.
<i>Change Match</i>	This tool toggles a single match on or off by picking the residue position on the sequence table or the molecule.
<i>Match Options</i>	This tool displays the Match Option dialog box. You can change the different match variables and cutoffs.
<i>Superpose Matched Residues...</i>	This tool superposes the matched residues of the active molecules. If a target molecule has not been selected, then the first active molecule is used.
<i>Save MSF...</i>	This tool saves the superposed molecule coordinates to their respective MSFs. It activates the standard MSF saving options.
<i>Reread MSF...</i>	This tool rereads the last saved version of the active molecules (MSFs) and restores the coordinates. This rejects any superposed coordinates that were not saved.
<i>Options...</i>	This tool activates the Align and Superpose Options dialog box from which five additional options can be selected.
<i>Superposition Options:</i>	Atoms to Superpose: By default only the Cα atoms are superposed but alternatives are to superpose all main chain atoms or for you to enter a selection. Choose target molecule: During the superposition one molecule will remain stationary. By default the first selected molecule is this target molecule.

5.. Align and Superpose

Output transformation matrix: If this option is checked then after each superposition the rotation matrix and translation vector applied to each molecule is listed to the textport.

Move all atoms in molecule: By default this option is checked on. If it is switched off then you will be given the atom selection palette in order to select the atoms which will move during the superposition.

Gap Penalties

This option presents a dialog box which allows you to change the penalties assigned to creating a gap in automatic alignment. The different forms of penalty function are discussed above. The dialog box has options for you to select which forms are active and to change the penalty value for each form. There is also an option to change the maximum gap length. By default the alignment algorithm will not test alignment which involve inserting gaps greater than 40% of the sequence length (this limitation reduces calculation time) but if you are working with some exceptional sequences you may wish to change this.

Dot Plot Options

Number of residues in window: By default dot plots are drawn for a single window length of 11 residues - you can change this value to anything between 1 and large values such as 31. Analysis for more than one window length can be presented on the same plot. Multiple window length can be entered in the text input line; the values should be separated by spaces.

Show constraints on dot plot: By default any constraint between residues of the two plotted sequences are shown on the dot plot.

Normalize dot plot scores: By default the coloring of dot plots uses the normalized scores where the normalization has been done over the entire dot plot. If this option is not checked then the dot plot will be colored according to the absolute scores.

Dot Plot Color Ranges

This displays the Color Range dialog box from which you can edit the dot plot colors and cutoffs

Color by Homology

The coloring of molecules and sequences is controlled by the **Molecule Color** tool on the **Protein Utilities** palette. One option is to color by the homology between sequences. The colors and cutoff values used in this coloring scheme can be changed in this dialog box.

RMS Deviations...

The RMS deviation between the currently active structures are calculated and listed to textport. By default a single figure of RMS per pair of molecules is listed. If the active range is on then this tool is applied to only the residues in the active range. There are several options:

Calculate for matched residue only

To calculate an rms deviation for only a limited set of residues you should ensure those residues are matched using either the **Match Residues** or **Change Match** tool and then check this option.

<i>List RMS per residue</i>	An rms deviation per residue is listed to the textport.
<i>Plot RMS per residue</i>	The rms per residue is plotted to the sequence viewer.
<i>Atom selection</i>	By default the reported rms is for just the C α atoms but alternative selections are available.
<i>Plot Dendogram</i>	This tool is available if there is a dendogram plot currently displayed. The dendogram will be written to a PostScript format file which can be used to create a hardcopy plot.
<i>Finish</i>	This tool exits the palette If structures have been superposed but not the coordinates have not being saved then you will be prompted to save them.

The Constraints Palette

The constraints palette is activated when the **Constraint** tool on the Align and Superpose palette is picked. The palette is closed by repacking the **Constraint** tool or picking the **Exit Constraints** tool from the Constraint palette. The palette has tools to enable selection of constraints, to save and restore constraints in external files and to toggle on or off the use of constraints in alignment.

To define a constraint you must select one residue per sequence for two or more sequences. Constraints are shown on the sequence viewer as a thin blue line between the residues. Beware that this line might be obscured by the Match indicator if it is active. If the **Add Constraint** tool is active then the last picked residue is indicated on the sequence viewer by a blue triangle under the residue. Constraints are also shown on dot plots by a blue circle about the position corresponding to two constrained residues.

By default once a constraint is selected or read in it will be used in any subsequent automatic alignment but the constraints can be excluded from the alignment by deactivating the **Use in Auto Align** tool. If not all sequences are active when an automatic alignment is performed then only the constraints with residues in two or more active sequences will be used. Constraints are not saved automatically between QUANTA sessions so any constraints required in future should be saved to file.

Constraint Palette Tools

<i>Add Constraint</i>	If this tool is active then constraints can be selected by picking the appropriate residues in the sequence viewer or on the molecules. The selected residues are indicated by pale blue boxes. Only one residue per sequence should be selected; if a second residue is selected from the same sequence
-----------------------	--

5.. Align and Superpose

then there is a warning message and the option to use either the previous or the new residue. Once a residue has been selected from all the currently active sequences then the constraint is considered to be completely defined and is saved and the next residue pick is considered as the start of a new constraint. It is also possible to define a constraint between two sequences by picking a point on a dot plot for those two sequences. It may be helpful to increase the scale of the dot plot by using the full screen icon at the top right of the dot plot window or by using the Zoom Window tool on the dot plot pull down menu under Display.

Restart

If a mistake is made in selecting residues for constraints then this tool should be used to restart the selection for the last constraint.

Next

This tool should be used after all the required residues have been selected for the current constraint. The constraint does not need to have a residue selected for every sequence but should have a residue for at least two sequences. The next residue you pick will start the definition of the next constraint.

Delete Constraint

If this tool is active then picking any one residue in a constraint will delete that constraint. Constraints can also be selected by picking the dot plot.

Delete All

Deletes all constraints.

Use in Auto Align

The constraints will only be used if this tool is active. If the Constraint palette is closed the status of this tool is retained for all subsequence alignments.

List to Textport

List all current constraints to the textport. The information is organized with each constraint on one line and all the constrained residues in each sequence in a column under the sequence name. A * character indicates that the constraint does not apply to that sequence.

Save to File

Save the data to a file with the default extension .con. The information is organized with each constraint on one line and all the constrained residues in each sequence in a column. The sequence names are given at the top of the file.

Restore from File

A constraint file with default extension .con is opened and the constraints read. If the name of a sequence in the file does not correspond to any currently selected sequence then the information for that sequence will be ignored but so long as the constraint still has two or more residues in currently selected sequences it is read in. If sequence or MSF names have been changed since the constraint file was written it is possible to edit the file to update the names.

Exit Constraints

Close the Constraint palette. Note that if constraints are selected and the **Use in Auto Align** tool is active then the constraints will be used in any future alignment.

6. Create Homology Model

Overview

This utility enables you to copy coordinates from a known structure or structures to a sequence whose structure is unknown. This process creates a homology model for the sequence that can be further refined using other modeling tools in Protein Design and other QUANTA applications, such as Conformational Search's Loop Modeling.

This chapter describes:

- Copying homology
- Tools and options

For more information see:

Protein User's Reference

- Align and superpose
- Edit protein
- Model backbone
- Model side chain

Copying Homology

This utility generates the framework of a homology model by copying structure from homologues. The mainchain atom coordinates are copied directly from a known structure. Sidechain atom coordinates are copied as far as possible and any unresolved sidechain atom coordinates are built by regularization. You must decide the most appropriate homologues to copy coordinates from and the regions of the structure for which copying is valid (see the Protein tutorial).

The most efficient means to generate a model is to use the **Auto Match** tool in Align and Superpose to identify regions of sequence homology between the sequence of unknown structure and the known structure. Using a **Match Window** of say three or five residues (see under **Match Options**) will better identify homologous regions and exclude individual good matches. The **Copy Matched Residues** tool can then be used to copy coordinates for all the matched residues at once.

6.. Create Homology Model

It is important to observe regions where two consecutive residues are modeled on different proteins. The two residues may not join well or may be too far away to form a bond. When this is observed, other tools should then be used to further refine the structure.

After remodeling the protein backbone by copying coordinates from a homologue, it is advisable to re-model the sidechains. The **Auto Model** tool on the **Model Side Chains** palette does this (see Chapter 9) and, to simplify the procedure, the Create Homology utility automatically writes a selection file, `copy_side_sel.rsd`, which lists the remodeled residues.

Tools

When the Create Homology Model utility is selected from the **Protein Design** palette the Choose an Unknown Structure dialog box opens with a scrolling list of all the active molecules. You need to select one of the molecules as the unknown structure. This sets that molecule as the structure being modelled. When the palette is displayed, all active molecules are listed under the **COPY FROM** tool, with the unknown structure grayed out. The unknown structure should be an MSF and not a sequence. MSFs can be created from sequences using the tool in the **Protein Editor** utility.

Change Unknown Structure...

This option displays the Choose Unknown Structure dialog box with a scrolling list of all active molecules. The active molecule selected becomes the new unknown structure. This option allows you to change the selection of the unknown structure from the initial structure.

Select Copy Range...

This tool activates the **Pick Range** palette. You can then pick the residue range over which the copying will be performed. This is done by picking the first and last residues of the range from either the sequence table or the molecule. Within this utility, the **Active Range** tool on the **Utilities** palette also has the same function as **Select Copy Range**.

Copy Options...

This option displays the Copy Options dialog box, from which you select options for the visual display and mode of the copied coordinates.

Copy Range

This tool remains grayed out until a copy range has been selected on the unknown structure. Once this tool is active, selecting it copies the coordinates of the chosen known molecule to the unknown structure.

Copy Matched Residues

Copy the coordinates for all the residues currently matched. Matching residues are determined in the Align and Superpose utility by either an automated procedure to find homologous regions or by user selection. Matched residues are highlighted on the Sequence Viewer by a vertical yellow band. This highlighting is usually switched off outside the Align and Superpose module, but it is switched on again upon entering the Create Homology

Save MSF...

Model utility. If the **Select Copy Range tool** is active, then coordinates will only be copied for the matched residues within the selected range.

Reread MSF...

This tool saves the changes to the MSF of the unknown structure. The standard MSF saving dialogs are displayed.

Finish

This tool reads the last saved version of the MSF into QUANTA. Any changes not saved to the MSF are lost.

This tool exits the **Create Homology Modeling** palette and any changes remain active in memory. If the changes have not been saved, the standard save dialog boxes are displayed for saving the structural changes.

6.. Create Homology Model

7. Superpose Folding Motif

Overview

This utility superposes structures on the basis of their overall folding rather than requiring identifying homologous residues. Using this utility, protein structures with similar folding motifs, but possibly little other obvious homology, can be superposed.

A folding motif can be either a whole protein or a specific area of a protein. It is defined in terms of the α -helix and β -strand secondary structure elements, and the inter-element geometry, such as distances and angles.

This chapter describes:

- Superposing folding motifs
- Protein geometry
- Secondary structure representation
- Sequence alignment
- Tools and options
- Demonstration of using Superpose Motif

For more information see:

Protein User's Reference

- Align and Superpose
- Domain Analysis
- Superpose Motif Database

References

E. M. Mitchell, P. J. Artymiuk, D. W. Rice; P. Willett, "Use of techniques derived from graph theory to compare secondary structure motifs in proteins," *J. Mol. Biol.* **212** 151-166 (1989).

A. T. Brint; P. Willett, "Pharmacocophoric pattern matching in files of 3D chemical structures: comparison of geometric searching algorithms," *J. Mol. Graph* **5** 49-560 (1987).

G. D. Mulligan; D. G. Corneil, "Correction to Bierstone's Algorithm for Generating Cliques," *J. ACM* **19** 244-247 (1972).

C. A. Orengo, N. P. Brown; W.R. Taylor, "Fast Structure Alignment for Protein Databank Searching," *Proteins* **14** 139-167 (1992).

Superposing Folding Motifs

A simple example of the problem of superposing two similar, but not identical, structure motifs is shown in Figure 1.

The reference motif has a set of parallel β strands with an α -helix lying across the strands. The test motif includes a β -sheet of four strands, with the two central strands parallel and two helices, one on either side of the sheet. No consideration is made of the connectivity between the secondary structure elements.

The secondary structures are represented simply as vectors. They are labelled R1 to R4 for the reference structure and T1 to T6 for the test structure. In this example, there is no combination of elements in the test motif that reasonably matches all the elements in the reference motif. There are four possible ways in which three of the four elements in the reference motif could be matched to elements in the test motif.

For this example, it is possible to determine all possible matches by eye. However, for more complex examples, an efficient algorithm to test all possible combinations is needed.

In the automated matching algorithm each structure is analyzed separately. The geometric relationship (i.e., distances and angles) between all pairs of secondary structure elements are calculated. These geometric relationships are then used to determine whether a pair of elements in one structure might be equivalent to a pair of elements in the other structure. If the differences in distances and angles are not too great then the two pairs of elements might be equivalent.

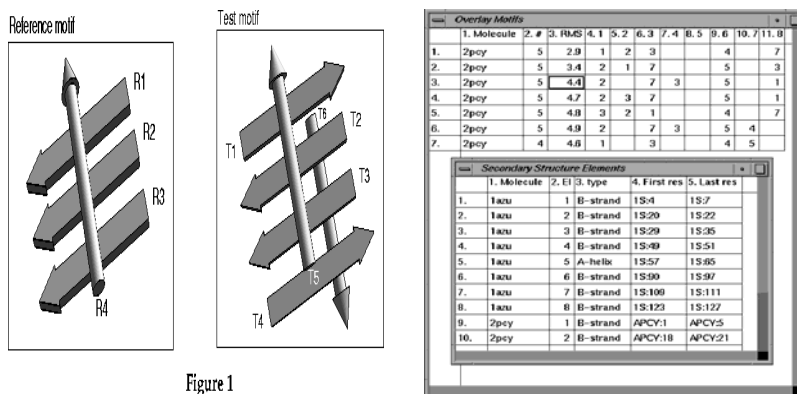


Figure 1

Table 1. Reference structures and matches

Reference structure	Match 1	Match 2	Match 3	Match 4
R1	T2	—	T3	—
R2	T3	T2	T2	T3
R3	—	T3	—	T2
R4	T5	T5	T6	T6

For the example in the figure the distance and angle between the two elements of the reference structure, R2 and R4, are similar to the distance and angle between the elements of the test structure, T2 and T5. However, they are dissimilar to the distance and angle between elements T1 and T5, since the strand T1 is running in the opposite direction.

The result of comparing all possible pairwise combinations of elements is recorded in a correspondence matrix which records either true or false for whether pairs of elements might be equivalent. A graph theory algorithm is used to analyze the correspondence matrix to find combinations of elements in the reference structure which will match the maximum number of elements in the test structure. Frequently there are several possible combinations of elements which give the same number of matches overall.

Given a set of possible matches between secondary structure elements the two structures can be superposed. This is done by applying a standard least squares superposition algorithm to the endpoints of the axial vectors of the matching elements. Combinations of matches which lead to poorly superposed elements (i.e., with a poor rms difference in the endpoint coordinates) can be removed from the list of matches.

Protein Geometry

The axial vector of a secondary structure element is defined as the principle moment of the C α atom co-ordinates. The endpoints of a vector are the projection of the terminal C α atoms onto the vector. The relationships between pairs of axial vectors are defined as:

Minimum distance

This is the minimum distance between the two axial vectors. When the closest point on one vector from another, is the vector endpoint, then the minimum distance is to that endpoint, rather than to any point on the line extended beyond the vector.

Average distance

This is the average of the distance between all the C α atoms in one secondary structure element and all the C α atoms in the other element.

7.. Superpose Folding Motif

Scalar Angle

This angle is derived from the inverse cosine of the scalar product of normalized vectors.

Tilt Angle/Interaxial Angle

This angle uses the definition given by Orengo* for representing the relative orientation of two axial vectors as two angles.

Same Domain

If the protein structures have been analyzed into domains then to satisfy “same domain” criteria pairs of secondary structure elements should have the same relationship “in same domain” or “not in same domain”.

Same Connectivity

The fold matching algorithm does not inherently require that matched secondary structure elements are in the same order along the protein chain but this requirement can be set.

Secondary Structure Representation

Secondary structure elements are represented by axial vectors. The **Secondary Structure tool** from the **Protein Utilities** palette will toggle the display of secondary structure vectors. In this utility it is recommended that the secondary structure vectors are displayed, but the molecule visibility be switched off.

To superpose only a limited fragment of a protein some secondary structure elements can be made inactive. The activity is toggled using the **Change Activity** tools on the palette. Secondary structure vectors that are active are represented by vectors with fat lines; and those that are inactive are represented by vectors with thin lines.

Sequence Alignment

If structures have been matched with the criterion that elements have the same connectivity, it is reasonable to attempt to find an optimal alignment of the protein sequence. The structures are superposed, and an alignment is performed to minimize the distance between C α atoms in aligned residues.

Tools and Options

This tools on this palette can be grouped interfere main functions: selecting the active secondary structure elements, matching the secondary structure,

*“Fast Structure Alignment for Protein Databank Searching”, C.A. Orengo, N.P. Brown & W.R. Taylor (1992), Proteins 14 139-167.

reviewing the matches and superposing and aligning structures based on one selected match.

Select Active Secondary Structure

Change Active

All secondary structure elements in a molecule are used by default. However, elements can be made inactive or unselected using one of the selection tools.

When **Change Active** is picked the **Pick Element**, **Pick Element Range**, and **Pick Domain** tools are ungrayed. Only one of these tools can be used at a time.

Pick Element

This option is used to toggle on and off the activity of a single secondary structure. The activity is toggled by picking an atom or residue. This is done either on the structure or sequence table, of a secondary structure element.

Pick Element Range

This option is used to select the activity of any two secondary structure elements. Either two atoms in the molecule or two residues in the sequence table. All elements within the selected range of the two elements are toggled on or off.

Pick Domain

If domain analysis has been performed on the molecule, this option selects one of the assigned domains. The activity is toggled by picking an atom or residue from the sequence table of a selected domain. If domains are unassigned, any pick selects a complete segment.

Match Secondary Structure

Overlay Motifs

When this tool is selected, all possible overlays for the two active molecules are calculated and the resulting vectors displayed. In the legend area, each overlay is numbered and listed, along with RMS difference after superposing the secondary structure vectors. Two tables are also displayed that show information about the motif overlays and secondary structures.

7.. Superpose Folding Motif

Overlay Motifs											
	1. Molecule	2. #	3. RMS	4. 1	5. 2	6. 3	7. 4	8. 5	9. 6	10. 7	11. 8
1.	2pcy	5	2.9	1	2	3			4		7
2.	2pcy	5	3.4	2	1	7			5		3
3.	2pcy	5	4.4	2		7	3		5		1
4.	2pcy	5	4.7	2	3	7			5		1
5.	2pcy	5	4.8	3	2	1			4		7
6.	2pcy	5	4.9	2		7	3		5	4	
7.	2pcy	4	4.6	1		3			4	5	

Secondary Structure Elements						
	1. Molecule	2. El	3. type	4. First res	5. Last res	
1.	1azu	1	B-strand	1S:4	1S:7	
2.	1azu	2	B-strand	1S:20	1S:22	
3.	1azu	3	B-strand	1S:29	1S:35	
4.	1azu	4	B-strand	1S:49	1S:51	
5.	1azu	5	A-helix	1S:57	1S:65	
6.	1azu	6	B-strand	1S:90	1S:97	
7.	1azu	7	B-strand	1S:109	1S:111	
8.	1azu	8	B-strand	1S:123	1S:127	
9.	2pcy	1	B-strand	APCY:1	APCY:5	
10.	2pcy	2	B-strand	APCY:18	APCY:21	

Motif Tables

These tables display information about the calculated motifs and secondary structures for each molecule.

The Overlay Motifs table displays numerical information on each of the possible overlays. Columns are, from left to right: the overlay number, the second molecule name, the number of elements matched, the RMS difference of the superposition. All subsequent columns identify elements that are matched. For example, column 6.3, row three has the number seven. This indicates that the third element in 2pcy was matched with the seventh element in 1azu.

The secondary structure elements table displays information on each secondary type in both active molecule. Columns are, from left to right: molecule name, element number, secondary structure type, the ID of the first

Motif Superposition Options

residue in the secondary structure, and the ID of the last residue in the secondary structure.

This tool displays the Motif Superposition dialog box that contains variables used in matching secondary structure elements and the match cut-offs. These variables set the minimum criteria for structures to be considered matched.

Number of matched elements

Only overlays which match a minimum number of secondary structure elements will be reported in the Motif Table and displayed. There are two alternative means to define the minimum.

- Minimum number of elements that are matched. This is an absolute integer value.
- Fraction of number of search elements. This is the fraction of the total number of elements in the first or search structure. This is a value between 0.0 and 1.0.

Superposition RMS difference

This is the root mean square (rms) difference in the coordinates of the ends of the axial vectors after superposition. This can be used as a test for the similarity of the position, orientation, and length of the vectors and if a match results in a poor overlay then it will be removed from the list of matches.

Individual Secondary Structure

The individual matched elements should satisfy the following criteria:

- **Same Secondary Structure Type:** The matched elements should be the same secondary structure type. By default this criteria is on.
- **Similar Element Length:** The matched secondary structure elements should be of similar residue length. The maximum number of residues difference is indicated, the default is 4.

Geometric Criteria of Secondary Structures

These are criteria for a pair of secondary structure elements in one structure to be considered similar to a pair of secondary structure elements in the other structure. Most of these criteria relate to the distances and angles between the pairs of elements.

- **Same sequence order:** By default elements do not need to have the same sequence relation, or connectivity, but this criteria will be applied if this options checked.
- **Minimum separation** The minimum distance between the axial vectors, and the average of all the C α atoms in both elements. The difference between matched pairs of elements should, by default, be less than 5 Å.
- **Average separation:** the average distance between all the C α atoms in one element and all of the C α atoms in the other element. The difference between matched pairs of elements should, by default, be less than 5 Å.

7.. Superpose Folding Motif

- **Inter-vector angle, Interacial angle, and Interacial tilt:** These are three possible means of measuring the angle between elements. A maximum accepted difference between matched pairs of elements is, by default, 40°.
- **Similar loop length:** For two consecutive, connected elements matches can be limited to elements connected by loops of similar residue length. This is useful in searching for a motif of just 2 or 3 consecutive elements.
- **Segment relationship:** For a pairs of elements to be matched they must have the same relationship were there are two possible relationships: both in the same segment or in different segments.

Reviewing the Matches

All Overlays

This tool is inactive until the **Overlay Motif** option is selected and calculated. It displays all of the calculated overlays in the viewing area and lists them with their rms values in the legend and in the textport.

Next Overlay

This tools is inactive until the **Overlay Motifs** tool is used and there is more than one match. This tool steps forward displaying each overlay.

Previous Overlay

This tool is inactive until **Overlay Motifs** tool is used and there is more than one match. This tool steps backward displaying each overlay.

Select Overlay

This tool is inactive until **Overlay Motif** tool is used. It presents you with a list from which to select one or more overlays to be displayed.

Clear Display

This tool is inactive unless **Overlay Motifs** tool is used. It removes the display of overlays and masks the browse tools.

Superpose and Align Molecules

Superpose Molecule

This is inactive unless one match is selected by the browse tools. Superpose the molecule co-ordinates on the basis of the displayed match. If the molecules are invisible then make them visible.

Reread MSF

This tool rereads the MSF and restores the atomic co-ordinates.

Save to MSF

This tool saves the current atomic co-ordinates to the MSF.

Align Sequence

When one overlay is selected this tool aligns the sequence based on minimizing the distance between C α atoms. The result of this alignment probably is only meaningful if structures are matched with elements in the same order.

Undo Alignment

This discards the current alignment.

Match Close Residues

When sequences are aligned, this tool indicates which pairs of aligned residues are close by placing yellow bars on the sequence viewer. This is similar to the **Match Residues** option on the Align and Superpose palette. The cutoff criterion for close residues is, by default, 2.5 Å.

Finish

Exit the **Superpose Folding Motif** palette. If molecules have been superposed but not saved, you are prompted to save coordinates to the MSF.

Demonstration of Using Superpose Motif

The following exercise demonstrates how to use the Superpose Motif palette. The active structures used in this example are 1azu and 2pcy.

1. **From the Molecule table, toggle the activity on and visibility off for structures 1azu and 2pcy.**
2. **From the Protein Utilities Menu, toggle on the tools Secondary Structure and Legend. Next, select the option Molecule Colors, and from the Molecule Colors dialog box.**

Select the options:

Color Mode

Secondary Structure

Color non-carbon atoms by element type color

Select Atoms to Display

Alpha Carbon atom trace

Click **OK** and the display and legend are updated, reflecting the selected changes.

3. **From the Protein Design Menu, select the utility Superpose Folding Motif.**
4. **From the Superpose Folding palette, select the tool:**

Overlay Motifs

The overlays are calculated, using the default **Motif** options, and displayed in the viewing area. The motif tables, Overlay Motifs and Secondary Structures Elements, are displayed and the browse tools are unmasked and activated. The legend lists all the overlays along with their RMS value.

5. **Select the browse tool:**

7.. Superpose Folding Motif

Next Overlay

The first overlay is displayed in the viewing area and legend. Click on the tool to step forward through the overlays, or, to step backwards, use the tool;

Previous Overlay

6. View the Overlay Motif Table. The calculations resulted in seven possible overlays, and these are listed in order of increasing rms value.

7. Select the tool Select Overlay(s)...

The Display Selected Fragments dialog box is displayed. Pick from the scrolling list overlay 1.

8. Select the option:

Match Close Residues

and the tools **Superpose Molecule** and **Align Sequence** are automatically selected.

The structure 2pcy is superposed on the 1azu molecule; the 2pcy sequence is aligned to 1azu to minimize C α -C α distances; and matches between the two molecules are calculated and reported in the textport.

9. Select the option

Reread MSF

The molecule 2pcy is reread into the work area, and the coordinates are restored to the saved version of the MSF.

10. Select the option;

Finish

The Superpose Folding Motif utility closes.

8. Model Backbone

Overview

The protein modeling tools are divided into two utilities: Model Backbone for defining mainchain conformations, and Model Side Chains for defining sidechain conformations. They are closely interdependent. Therefore, when modeling, first examine possible mainchain conformations, and then review the sidechain conformation.

This chapter describes:

- Modeling the protein backbone
- Tools and options definitions

For more information see:

Protein User's Reference

- Model Side Chains
- Analyze Secondary Structure)
- Copy Coordinates
- Predict Secondary Structure
- Edit Protein

Modeling the Protein Backbone

Homology modeling is building a model of a protein of unknown structure based on a homologous known structure or structures. An initial model can be generated by:

- Using the Protein Editor to perform the necessary insert, delete, and mutate operations
- Copying coordinates to copy the conformation of a known protein onto the sequence of a protein being studied

Using either procedure usually results in a structure with some regions of uncertain conformation. The protein modeling tools in Protein Design are divided into two groups: one defines the main chain conformation, and the other defines sidechain conformations. While these two groups are closely interdependent, the dependency is not well understood. The best approach

to take in actual modeling, therefore, is to first consider all possible main chain conformations and then consider the side chain conformations.

The Model Backbone utility has several different tools for refining a structure: regularizing, building coordinates, folding residue ranges and fragment building using the fragment database. All of these tools work specifically on the protein backbone.

Regularizing Regions

All regularization in the Protein Design module uses an internal minimizer and the idealized geometry which is stored in the files \$HYD_LIB/protein_structure.gsd or the binary version of the same file \$HYD_LIB/protein_structure.bgsg (see Appendix B). The regularizer tool is available in several utilities where it might be most useful: for example in the Fragment Database utility, the join between the modeled region and the rest of the structure often has poor geometry but this can be improved by regularization.

Regularization is a means of cleaning up bad geometry in a model. It is an energy minimization procedure that takes into account geometry energy terms, such as bond length, angles, and torsion. By default, in regularization van der Waals energy terms are not considered, so this method will not remove bad interatomic contacts. It is possible to set regularization to consider van der Waals interactions. While this tool will improve the appearance of the model, it should be used sparingly.

Although judgement should be used, regularization is generally chosen to model undetermined regions in the following situations:

- The undefined residues are terminal residues.
- The region of undefined atoms is relatively short, for instance one to three residues in length.
- The undefined region does not have residues that can be used as an anchor.

Folding Residues

The protein conformation can be altered by changing the main chain dihedral angles. This option provides a quick means of alternating the conformation of range of residues to a user-specified pattern.

When the backbone dihedrals are changed one end of the protein chain can remain in the same position but the moving end of the protein chain could potentially move a significant distance thorough space. There are three

alternatives for determining which residues move: keeping the N-terminus fixed, keeping the C-terminus fixed, and retaining the overall average position. The large movement of one terminal is often not desired and can be avoided by using the option to break the protein chain. By default the break is made at the non-fixed end of the folded fragment but it can, optionally, be made at any point between the folding fragment and the chain terminal.

The alternative approach, to maintain the average position of the folded fragment, is done by doing a least squares superposition of the folded fragment onto the original C α coordinates.

Fragment Searching

The **Fragment** database searching finds fragments with appropriate geometries for modeling a small region. Once a fragment is selected, its conformation is copied to the model structure.

You must select a search template of three or more residues. These should be residues of reasonably known position either side of the uncertain region of the model structure. For example when searching for a fragment to model a loop region between two secondary structure elements you would pick the two terminal residues from each of the two secondary structure elements.

The database search protocol analyzes the inter-C α distances of the residues in the search template and searches the database for the same pattern of residues with similar inter-C α distances. The fragments with lowest differences in inter-C α distances are retrieved from the database and are displayed superposed over the search template. The RMS deviation for the least squares fit of the fragment over the search template is one of the parameters listed to the textport. The other listed parameter is the difference of the inter-C α distances calculated in the database search. You can review these fragments and can select one fragment to use to model the undefined region.

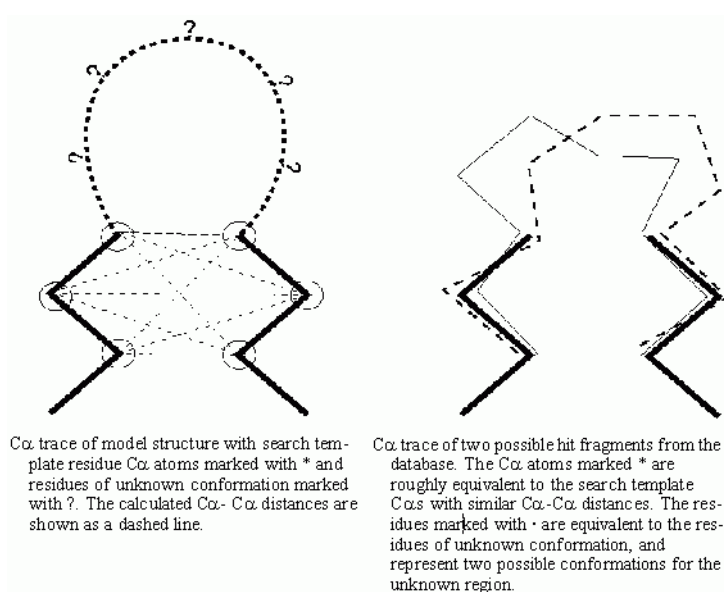
Useful criteria for choosing a fragment are:

- The fragment forms a good fit with the residues on either side of the undetermined region. The rms difference after the least squares fit is low.
- There are minimal close contacts to neighboring regions of the protein.
- The residues in the fragment are similar to those in the sequence of the model structure. If the modeled region includes glycine or proline residues, which have unusual main chain conformations, ensure they are given sensible conformations.

8.. Model Backbone

The fragment searching can also use the Bumps option. This takes retrieved fragments and fits them over the search template. The inter-atomic distances between the main chain, and C β atoms of the fragment and the neighboring residues are calculated. Those fragments with bad close contacts are then rejected. Since this procedure reduces the number of fragments finally selected, the initial database search will retrieve extra fragments.

The fragments retrieved and displayed within QUANTA come from a library of MSF files. If this library contains compressed files then the files need to be uncompressed before reading. To do this QUANTA will create a directory TMP_MSFLIB below your current working directory and copy the uncompressed MSF files to that directory. This directory can be deleted after you have completed this work.



Tools and Options

In the Model Backbone application, only one molecule can be active. If there is more than one active molecule when the application is entered, the first molecule is left active and the rest are set inactive. The active molecule can be changed by selecting a different molecule from the Molecule Management table.

Many of the tools in this utility are applied to a range of residues. The **Pick Range** tool on the **Protein Utilities** palette can be used to select and dese-

lect ranges. If you pick a tool with no range selected then the **Pick Range** palette will be made available for you to select a range. Note that the range remains selected until you deselect it.

<i>Regularize Region...</i>	This option prompts you to make a selection by displaying the Pick Range palette if an active range is not selected. The Select Active Range tool on the Utilities palette can also activate the Pick Range palette. If some of the atoms within the active range are undefined, the structure is built with idealized geometry. The regularization can include interatomic interactions (so it is really an energy minimization) which can be restricted to those between atoms in the selected region or it can include interactions with the neighboring atoms. The mode of action can be changed in the Regularization Options dialog box.
<i>Regularization Options...</i>	Offers options on whether the interatomic interactions are taken into account in regularization.
<i>Build Coordinates...</i>	This tool generates a structure for any undefined atoms in the structure according to the idealized geometry in the \$HYD_LIB/protein_structure.gsd file but does not perform any optimization of conformation. The Pick Range palette is displayed to select a range of residues.
<i>Apply Conformation...</i>	This option alters the protein conformation by changing the main chain dihedral angles. This option presents the Fold Protein Main Chain dialog box, from which a new fold type can be selected, and which determines the extent the structure is moved when folded.
<i>Copy fold from another residue range</i>	Displays the Pick Range palette from which a range of residues can be selected. These residues need not be in the current active molecule. If the number of residues within the selected range is not equal to the number of residues in the folding fragment, then the appropriate number of residues after the first residue in the selected range will be used.
<i>Fold to assigned secondary structure type</i>	Analyze Secondary Structure and Predict Secondary Structure applications are used to assign a secondary structure type to a residue. If the residue is assigned alpha helix or strand, then it is folded to the idealized conformation for that secondary structure type. The dihedral values are taken from the main chain fold data in the file \$HYD_LIB/protein_param.dat.
<i>User specified phi, psi and omega</i>	A dialog box with the ϕ , ψ and ω of all the residues in the active range is displayed. You can change the values.
<i>Fold in regular structure</i>	Displays a scrolling list from which to make a choice from a library of idealized secondary structure types. This library contains both repeating conformations, such as helix or strand, and structures of a finite number of residues, such as β -turn. These conformations will be applied to all residues in the active range, except for structures such as β -turns that are applied to the appropriate number of residues in the active range. The data for this library is stored in the file \$HYD_LIB/

8.. Model Backbone

Position folded fragment

protein_param.dat after the keyword **FOLD**. This file can be appended by the user.

Select either the **Fix N-terminus end of fragment**, **Fix C-terminus end of fragment**, or **Retain** average position of the fragment.

Carry connected residues

By default when the backbone torsions are changed the whole of the non-fixed terminal of the segment beyond the rotated bonds will move with the rotation. If you do not want this to happen you can opt not to carry the terminal or to carry only a limited range of residues which you will then be prompted to select.

Spin search side chain conformations

Toggles the option to spin search the side chains to find their optimal conformation.

Search Fragment Database...

This option displays a palette with a set of searching and browsing tools for modeling fragments.

The inter-C α distance matrices for a representative set of proteins is saved in the file \$QNT_ROOT/dmatrix/dmfile. You can create your own versions of the file or access an alternative file by changing the file name from the **Options** tool. The MSF files are read from a library directory that you can set with **Options** tool.

The currently selected residues are indicated by a red cross on the C α atom position and red boxes around residues on the **Sequence Viewer**.

Initially, all fragments are displayed superposed on the template residues. They are color coded on the structure and on the legend displayed on the right side of the screen. The legend gives the name of the protein from which the fragment is taken, its distance fit, and the RMS difference in C α atom position when the fragment is superposed on the template.

After remodelling the protein backbone by copying coordinates from a database fragment it is advisable to remodel the side chains. The **Auto Model** tool on the **Model Side Chains** palette will do this (see Chapter 9) and, to simplify the procedure the Fragment Database utility, automatically writes a selection file, fragment_side_sel.rsd, which lists the remodelled residues.

List Proteins

This option list to the textport the proteins in the currently active C α distance matrix file.

Pick Alpha Carbon Range

This option selects template residues by picking the first and last residue in a range.

Pick Alpha Carbon

This option templates residues by picking each individual residue.

Undo Last

This option deletes the last selection.

Undo All

This option deletes all selections.

<i>Search Database</i>	This option searches the fragment database by C α distance for matches to the currently selected residues.
<i>with Bumps</i>	If this option is active, any database search is followed by Bumps checking before the optimal retrieved fragments are displayed.
<i>Display All Fragments</i>	This option displays all the fragments.
<i>Display Next</i>	This option displays the next fragment on the list and removes all others from the viewing area.
<i>Display Previous</i>	This option displays the previous fragment on the list and removes all others from the viewing area.
<i>Select Display...</i>	This option opens the Display Selected Fragments dialog box with all the fragments listed, allowing you to select one or more for display.
<i>List Residues</i>	When only one fragment is displayed this lists the residue name and ID of each residue in the fragment, and the corresponding residue in the active molecule.
<i>Accept Fragment</i>	This option is grayed unless only one fragment is displayed. It then copies the coordinates of the fragment onto the corresponding residues of the active molecule.
<i>Reject Fragment</i>	This option clears all fragments from the display
<i>Regularize Joins</i>	A short range of residues, by default two residues, either side of the joins between the inserted fragment and the rest of the structure are regularized to correct any poor bond lengths and angles. The range of residues which can move in the regularization can be changed under the Options tool.
<i>Options...</i>	This tool opens the Fragment Modeling Options dialog box which allows you to change the number of fragments displayed after a search, how fragments are displayed and the dmfile used.
<i>Undo Last</i>	This tool restores the atom coordinates to the previous state, undoing the last modeling operation.
<i>Finish</i>	This tool exits from the Model Backbone palette with any changes made retained in memory.

9. Model Side Chains

Overview

This utility contains tools for modeling the protein side chain conformations. It is assumed that the protein main chain has been determined using the Model Backbone utility.

When side chains are altered by either the **Mutate** tools in the Protein Editor or the **Copy** tools in Create Homology Model the conformation of the side chain is retained as much as possible from the original structure and generally in homology modeling it is best to retain as much as possible from the homolog but for residues for which there is no homology evidence on which to base the side chain conformation this utility provides rotamer library and energetics tools to best fit the side chain.

This chapter describes:

- Modeling Sidechains
- Tools and Options

For more information see:

Protein User's Reference

- Create Homology Model
- Edit Protein
- Model Backbone

Modeling Sidechains

Sidechains should not overlap with neighboring residues so this utility includes tools to indicate close contacts and to perform energy minimization which will attempt to eliminate close contacts. But minimization will only find local minima and the rotamer and spin tools should be used to search through conformational space. Side chain modeling can be performed in a manual mode, analyzing each side chain individually. Alternatively, an automatic mode will allow you to select multiple residues which will all be fitted using rotamer libraries and minimization.

Automatic Side Chain Modeling

There is a tool to perform modeling of any number of selected side chains using one selected rotamer library and optionally using regularization. For each side chain, all rotamer conformations will be tested and the one with

least close contacts selected. The energy minimization will then find the best local conformation. The minimization does take account of van der Waals interactions.

This tool can be used after the protein backbone has been significantly remodeled; for example in the Create Homology Model utility, in Model Backbone, Fragment Database and in the Protein Editor. In all of these utilities, when a section of backbone is remodeled the affected residues are listed to a QUANTA selection file and this file can be used to select the residues for automatic side chain modelling. The selection files generated automatically by the Create Homology Model, Fragment Database and Protein Editor utilities are called `copy_side_sel.rsd`, `fragment_side_sel.rsd` and `edit_side_sel.rsd`, respectively. A new file is created on entering the utility and any existing file is overwritten so if you wish to save one of these selection files you must move it to a new file name.

The selection of residues for automatic side chain modeling is done through a standard selection palette which has the option to read a selection from a file, **Read Selection-Commands** on the **Selection Utilities** palette.

Close Contacts

Close contacts can be displayed using the **Display Contacts** tool. The criteria for determining close contacts is basically the same as used in the **Bumps** tool in fragment searching in the Model Backbone utility. All atoms closer than the specified bump cut-off distance are flagged. The default is set at 3.0 Å. If the structure includes hydrogens, then the bump cutoff is further reduced.

Rotamers

There are several analyses of the protein database that classify commonly occurring conformations for each residue type. These classifications are called rotamers. The Protein Design application uses three of these analyses in modeling sidechains. The three rotamer libraries are Ponder and Richards; Sutcliffe; and Dunbrack and Karplus. These three rotamer libraries are based on different analyses of the side chain dependence on the backbone conformation. See the Protein Health chapter for more information.

The Ponders and Richards analysis ignores the main chain conformation of a residue. The Sutcliffe analysis specifies whether each rotamer is for a helix, strand, or any main chain conformation.

Dunbrack and Karplus base their analysis on the side chain conformation as a function of the main chain ϕ and ψ angle. A statistical analysis for each

residue type groups together residues with ϕ/ψ values within a given range of positions on a two-dimensional grid. For each group of residues of similar main chain conformation, the number of occurrences of each possible side chain rotamer is counted.

The possible sidechain rotamer conformations for chi χ_1 and χ_2 dihedrals are defined as:

gauche +torsion range centered on $+60^\circ$

gauche -torsion range centered on -60°

*trans*torsion range centered on 180°

The subsequent dihedrals for longer sidechains are ignored.

When the current residue is modeled, the library is searched for the data for the ϕ/ψ grid point closest to the ϕ/ψ of the current residue. All rotamers with one or more occurrences for that grid point are considered. The χ values of the current residue are set to the ideal values for each rotamer.

Spinning Side Chains

The **Spin** tool scans the conformation space for the side chain by incrementing the dihedral by some fixed amount. The conformation is then tested for close contacts with neighboring residues. When a conformation without close contacts is found, it is displayed. For the longer side chains with two or more variable torsions, the search works by rapidly rotating the most remote bond from the main chain.

The default spin increment is 30° . When the initial conformation has no close contacts, the spin algorithm assumes a minimum energy well and ignores all acceptable conformations until a conformation with close contacts is found. After the spin search has covered the whole conformational space, the side chain is returned to its initial conformation. If none of the conformations are without close contacts, the spin search is repeated with the bump cutoff distance decreased, allowing marginally closer contacts.

Tools and Options

There are two modes for using this utility — the automated mode allows you to select all the residues of interest and then goes through them all automatically finding the rotamer which fits with fewest bad contacts and then, optionally, minimizing the residue.

9.. Model Side Chains

In the manual mode, the residue whose sidechain is currently being modeled is designated the current residue. There are tools which can either pick the current residue or step forward or backward through the sequence.

There are a series of tools for modeling the current residue. Only one tool can be active at a time, while the remaining tools are grayed out. When more than one conformation is possible, as with the rotamer libraries, the **Spin** and **Copy Homologous** tools, then the **Next Conformation** tool can be used to step through the possible models.

At the bottom left of the display, the initial and current torsions for the current residue side chain are displayed. At the bottom of the display, a text line reports the identity of the current residue and which modeling method was used to generate the current model. The text line also gives a conformation number for those methods, such as rotamer libraries, and spin which generate multiple models. For the rotamer libraries, the percentage of the side chains observed in this conformation is reported, and for the Karplus rotamers, the main chain conformation is also reported.

As you step through the display of multiple conformations for the rotamer libraries or for spinning, the number of close contacts and possible hydrogen bonds is reported in the textport.

Auto Model

The **Residue Selection** palette is presented so you can select any number of residues. You then have the option of which rotamer library to use and whether or not to minimize. The procedure runs through all the selected residues, finding the rotamer with the fewest bad contacts and then optionally refining that conformation.

Current Residue

This tool allows you to pick the next residue, either in the molecule or in the sequence viewer, that will become the current residue.

Next Residue

This tool selects the next residue in the sequence that becomes the current residue.

Previous Residue

This tool selects the previous residue in the sequence that becomes the current residue.

Display All

This tool toggles on the display of all atoms.

Display Sphere

When modeling side chains, it is often useful to limit the display to a sphere around the current residue. If multiple molecules are displayed, then for each non-active molecule the display sphere is taken around the residue equivalent to the current residue. It is helpful to have any homologous proteins correctly superposed over the active molecule. By default, the display sphere is 6 angstrom, but this can be altered with the **Options** tool.

Display Current

This tool displays only the current residue and any equivalent residues in non-active molecules.

<i>Display Contacts</i>	This tool displays close contacts between the current residue and neighboring residues in the active molecule. As the side chain conformation is changed, the contacts are updated.
<i>Build Side Chains</i>	This tool activates the Selection palettes that allow you to select residues. If any atom in these selected residues has undefined coordinates, then templates for idealized side chain geometry found in \$HYD_LIB/protein_structure.gsd are used to generate the atom coordinates.
<i>Reset</i>	This tool restores the current residue to its initial conformation. If several residues have been modeled, they all are restored to their initial conformation by the ReRead MSF tool. It is advisable to save the modeling results frequently using the Save to MSF tool.
<i>Spin Residue</i>	This tool increments the side chain torsions by 30° or by the value set by Options tool, until it finds a conformation with no close contacts or a minimum number of close contacts. The Next tool steps to the next conformation with no contacts.
<i>Manually Rotate</i>	This tool activates a pseudo-dial set for rotating bonds.
<i>Copy Homologous</i>	This tool copies the side chain conformation from the equivalent residue—the aligned residue in the sequence viewer—to the current residue. If there is more than one equivalent residues, the Next Conformation tool can be used to step through displaying each of them. The actual number of torsions copied for all possible residue type pairs is defined in the Equivalent Torsion Lookup Table in the file \$HYD_LIB/protein_param.dat. Any remaining torsions in the current residue side chain retain their previous value.
<i>Ponders Rotamer</i>	This tool sets the current residue to the optimal rotamer of the Ponders and Richards rotamer library. If there is more than one rotamer, the Next Conformation tool can be used to step through them.
<i>Sutcliffe Rotamer</i>	This tool sets the current residue to the optimal rotamer of the Sutcliffe rotamer library. If there is more than one rotamer, the Next Conformation tool can be used to step through them.
<i>Karplus Rotamer</i>	This tool sets the current residue to the optimal rotamer of the Dunbrack and Karplus rotamer library. If there is more than one rotamer, the Next Conformation tool can be used to step through them.
<i>User Defined...</i>	This tool provides a dialog box to changes the required torsions.
<i>Minimize</i>	Energy minimization is performed for the individual residue.
<i>Options...</i>	This tool displays the Side Chain Modeling Options dialog box from which you can change default variables. The options are described in the following section.
<i>Bump cutoff</i>	This specifies, in angstroms, the minimum allowed distance between atoms below which Close Contact is displayed.

9.. Model Side Chains

Hydrogen Bond Cutoff

This specifies, in angstrom, the maximum distance between hydrogen bond donor/acceptor pairs used in analyzing the number of possible hydrogen bonds when comparing rotamer conformations.

Spin Increment

This specifies the increment of the side chain torsion (in degrees) used in the **Spin Residue**.

Radius of display sphere

This specifies in angstroms the radius when **Display Sphere** tool is used.

Protein parameter data file

This specifies the file containing side chain modeling and other protein modeling parameters.

Harvard rotamer data file

This specifies the file containing data for the Karplus rotamer library.

Next Conformation

When there are multiple possible conformations for the rotamer libraries, this tool enables you to step through them.

Save to MSF

Save the current atomic coordinates to MSF

ReRead MSF

Restore the atomic coordinates from the MSF file.

Finish

Exit the Model Side Chain utility.

10. Analyze Secondary Structure

Overview

This utility provides tools that assign secondary structures to proteins. By defining the secondary structure of a molecule, the shape of the molecule can be visualized. This is a two-step procedure; hydrogens bonds are calculated and then secondary structures are assigned based on the hydrogen bond patterns.

This chapter describes:

- Analyzing Secondary Structures
- Hydrogen Bond Calculations
- Secondary Structure Assignment
- Tools and Options Definitions

References

W. Kabsch and C. Sander, Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features, *Biopolymers*, **22**, 2577 (1983)

Jane S. Richardson, The Anatomy and Taxonomy of Protein Structure, *Advances in Protein Chemistry*, **34**, 167-218 (1981).

Richardson, J.S., Getzoff, D.C., and Richardson, D.C., Proceedings of the National Academy of Science, USA **75**, 2574-2578 (1978).

analyze
2nd struc

Analyzing Secondary Structures

Hydrogen Bond Calculations

In the method of Kabsch and Sanders hydrogen bonds are determined by an energy calculation. Minimum or maximum energy values for electrostatic interactions among the amide NH and CO are used to assign hydrogen bonds. This calculation requires the coordinates of the amide hydrogen atoms. If these coordinates are missing from the MSF, hydrogen bonds are calculated according to amide N-O cutoff distances and C-O-N angles.

The calculation uses the formula in Kabsch and Sander:*

10.. Analyze Secondary Structure

$$E = q_1 \times q_2 \left(\frac{1}{r(\text{ON})} + \frac{1}{r(\text{CH})} - \frac{1}{r(\text{OH})} - \frac{1}{r(\text{CN})} \right) \times f \quad \text{Eq. 1}$$

Where the coefficient values are:

$$q_1 \equiv 0.42e \quad q_2 \equiv 0.20e \quad f \equiv 332 \quad \text{Eq. 2}$$

and

$$E < -0.5 \quad \text{Eq. 3}$$

for a hydrogen bond.

Secondary Structure Assignment

The following conventions are used below in the secondary structure definitions:

$C\alpha$ torsion for the i th residue is defined by $C\alpha(i-1)$, $C\alpha(i)$, $C\alpha(i+1)$, and $C\alpha(i+2)$

$\text{hbond}(i, j)$ is a hydrogen bond between the amide O of residue i and the amide N-H of residue j

The following types of secondary structures are recognized:

As defined by Kabsch and Sanders using hydrogen bonding patterns

- α helix: Two or more consecutive 4-turns.
- β strand: **Two consecutive residues in one strand must have hydrogen bond bridges to two consecutive residues in another strand. A hydrogen bond bridge exists for a pair of** residues, i and j , if there are two hydrogen bonds near residues i and j in one of the following combinations: $\text{hbond}(i-1, j)$ and $\text{hbond}(j, i+1)$, $\text{hbond}(j-1, i)$ and $\text{hbond}(i, j+1)$, $\text{hbond}(i, j)$ and $\text{hbond}(j, i)$, or $\text{hbond}(i-1, j+1)$ and $\text{hbond}(j-1, i+1)$.
- 3-turn: Residues $i+1$ and $i+2$ are a 3-turn if $\text{hbond}(i, i+3)$ exists.
- 4-turn: Residues $i+1$, $i+2$, and $i+3$ are a 4-turn if the hydrogen bond $\text{hbond}(i, i+4)$ exists.
- 5-turn: Residues $i+1$, $i+2$, $i+3$, and $i+4$ are a 5-turn if $\text{hbond}(i, i+5)$ exists.

*W. Kabsch and C. Sander, Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features, *Biopolymers*, **22**, 2577 (1983)

*β -bulges described by Jane Richardson**

Defined by $C\alpha$ pseudo-torsion angles

- β bulges: In adjacent β -strands, between two β -strand type hydrogen bonds, two residues (called position 1 and 2) on one strand are opposite one residue (called position X) on the other strand.[†]
- **Extended Conformation:** $C\alpha$ torsion between -180° and -130° ; or $C\alpha$ torsion between -130° and -100° (the latter occurs when $\phi \leq 100^\circ$ and $\psi > 100^\circ$). This conformation is indicative of β strands
- **Folded Conformation:** This is defined as two or more consecutive residues have $C\alpha$ torsions between 30° and 70° . For $C\alpha$ -only structures this classification is a strong indication of α -helix.

Tools and Options

Calculate Hydrogen Bonds

This utility calculates the secondary structures for the active MSFs. The tools and options in this palette are primarily designed to recalculate the assigned secondary structure using user-specified variables.

Display Hydrogen Bonds

This tool calculates the hydrogen bonds between mainchain atoms for active molecules.

Add Hydrogen Bonds

This tool toggles the display of hydrogen bonds. It can only be used after the hydrogen bonds have been calculated using **Calculate Hydrogen Bonds**, or hydrogen bonds have been added using **Add Hydrogen Bond**.

Delete Hydrogen Bonds

This tool adds a hydrogen bond between main chain donor or acceptor pairs. It will not add a hydrogen bond where one has already been calculated and is currently displayed, or between wrongly matched pairs of atoms. Two atoms must be picked. The tool remains active until it is deselected.

Calculate Secondary Structure

This tool deletes a hydrogen bond that has been calculated and displayed. Two atoms must be picked. The tool remains active until it is deselected.

Use Alpha Carbons Only

This tool recalculates the secondary structures based on changes you may have made. Otherwise the default values are reread.

Assign Secondary Structure

Sets the mode of action using only the $C\alpha$ torsion angles in the analysis of the secondary structure.

This tool changes the secondary structure assignments for single residues or range of residues. The mode of the residue selection is determined by the next two tools.

*Jane S. Richardson, The Anatomy and Taxonomy of Protein Structure, *Advances in Protein Chemistry*, **34** 167-218 (1981).

[†]Richardson, J.S., Getzoff, D.C., and Richardson, D.C., *Proceedings of the National Academy of Science, USA* **75**, 2574-2578 (1978).

10.. Analyze Secondary Structure

The default for editing secondary structures is to edit single residues. These residues can be picked either from the sequence table or from the display.

Pick Residue

This option selects single residues for editing secondary structures. This is the default mode for **Assign Secondary Structure**.

Pick Residue Range

This option activates the **Pick Range** palette, which enables a range of residues to be selected for editing.

Calculation Options...

This tool displays the Secondary Structure Analysis Parameters dialog box for redefining the different cutoff parameters.

The dialog box shows the default options.

Color Options...

This tool displays the Color Secondary Structure dialog box from which you can change the default color designation for secondary structures.

Write to File

This tool lists the secondary structure information to a file with the name *MSFname_secstr.out*. The column headings for the output are:

- Residue: Segment ID, residue number and residue name
- Ca tors: C α torsion angle
- Phi: ϕ angle
- Psi: ψ angle
- Hbond acceptor: Segment ID and residue number of hydrogen bond acceptors
- Sec type: Type of secondary structure assigned

List to Textport

This tool lists the information about the secondary structure assignments to the textport. This information is the same as described in the previous tool.

Save to MSF

This tool displays Secondary Structure Title dialog box for saving the secondary structure information to the MSF as extra information.

The dialog prompts for a title and uses the title default for the first analysis. Each change that is saved can be given a title. If you choose to overwrite existing data, the standard MSF Save Option dialog box is displayed.

Read from MSF

This tool displays the **Secondary Structure Title** dialog box from which previously saved secondary structure information can be read. All saved titles are listed.

Finish

This tool exits the **Analyze Secondary Structure** palette.

11. Calculate Accessibility

Overview

This utility contains tools to calculate solvent accessibility and to calculate contact areas between two molecules or two regions of the same molecule. There are several different coloring schemes to indicate the accessibility, or contact area, of atoms, whole residues or residue side chains.

This chapter describes

- Calculating Accessibility
- Tools and Options

For more information see:

Protein User's Reference

- Align and Superpose
- Domain Analysis
- Motif Database

Reference

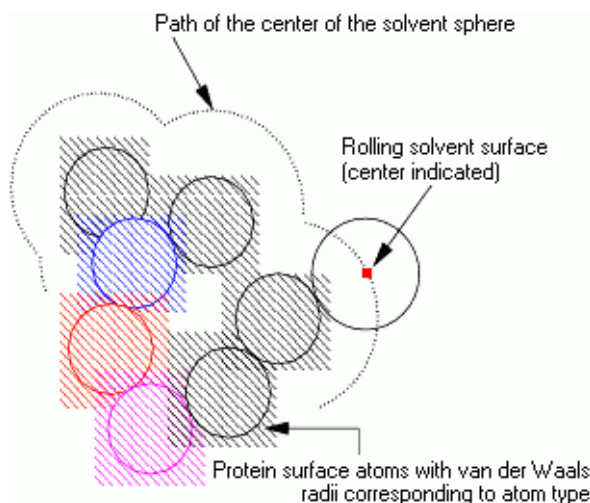
B. Lee and F. M. Richards, "The Interpretation of Protein Structures: Estimation of Static Accessibility" *J. Mol. Biol.* 55, 379-400 (1971).

Calculating Accessibility

The solvent-accessible surface area, or accessibility, of an atom is the surface area of the atom that is exposed to solvent. The residue accessibility is the sum of the accessibilities of the atoms in that residue. In studying proteins, the residue accessibility is a useful indicator to the residue's location, on the surface or in the core. One factor which should be taken into consideration in interpreting the data derived by this utility is that, because proteins are dynamic, even residues which are calculated to have very low solvent accessibility will be intermittently exposed to solvent.

Accessibility Calculations

In the conventional model for the accessibility calculation the solvent water molecule is represented by a sphere of radius 1.4 Å. A simple model for the calculation is of the solvent sphere being rolled over the surface of the molecule.



The path of the center of the solvent sphere is considered to be the accessible surface of the molecule. The distance from a protein atom center to the accessible surface is, at minimum, the van der Waals radius of the atom plus the solvent sphere radius. Because of the finite size of the solvent sphere, it cannot penetrate small interstitial spaces between atoms of the molecule surface. Hence the solvent surface appears smoother than the van der Waals surface.

Assigning a numerical value for accessibility requires a method of integration over the surface area. QUANTA uses the Lee and Richards method. It takes z-sections through the molecule, tracing a solvent accessible path around the molecule in each section. The length of the arc line is then calculated about each atom. This length is multiplied by the value of the z-spacing to give a rough estimate of area.

The z-spacing used is proportional to the sum of the solvent sphere radius and the minimum van der Waals atomic radius of the atoms in the molecule. The default proportionality constant, or z-spacing factor is 0.05.

When calculating accessibility for a protein it is necessary to exclude any associated solvent molecules and hydrogen atoms. Since accessibility calculations for a large protein can be slow, it is sometimes useful to perform the calculation for a selected set of atoms. However, a shell of neighboring

atoms must be included for the context of the calculation. These neighboring atoms are responsible for occluding some surfaces of the selected atom set. When setting up a calculation for a selected set of atoms, it is probably safest and simplest to do the calculation in the context of the whole molecule.

Contact Area

The contact area between two sets of atoms is a measure of their surface areas in contact. The two sets of atoms might be two molecules or two regions in one protein (e.g., two neighboring α -helices).

The definition of the contact area for the first set of atoms is the area that is accessible in the absence of the second set of atoms, but occluded in the presence of the second set of atoms. A similar definition applies for the contact area of the second set of atoms. The contact area of the two sets does not have to be the same, but normally it is similar.

To find these contact areas three accessibility calculations are done:

- Atom set 1 alone
- Atom set 2 alone
- Atom sets 1 and 2 together

The contact area for each atom is the difference between its accessibility in its own set and in both sets together.

Displaying Accessibility and Contact Areas

To display the accessibility or contact area on the molecule there are three alternative color coding schemes: atom, residue, or residue side chain accessibility.

The maximum possible accessibility of a residue side chain obviously depends on the residue type — bigger side chains have greater surface area. The fraction of the side chain accessible can be more meaningful than the absolute value. To calculate the fraction of the side chain which is accessible you also need to know the maximum possible side chain accessibility for each amino acid type.

Estimates for the maximum possible accessibility of each amino acid side chain are listed in the file, \$HYD_LIB/protein_param.dat, after the keyword ACCESS. These values were calculated for a fragment GLY-X-GLY. The backbone and the side chain of X were in extended conformation. There were no other atoms nearby to occlude the side chain, so the accessibilities of the side chain atoms were maximized.

11.. Calculate Accessibility

Since the maximum accessibility of a side chain depends on the side chain conformation, this fraction is just a rough guide and can even take values greater than 1. When calculating the accessibility, there is the option to calculate the maximum possible accessibility for each side chain in its current conformation. This is done by calculating the accessibility of the side chain in the absence of any neighboring residues. This value is then used in calculating the fractional accessibility of the residue.

Tools and Options

<i>Accessible Area...</i>	This option calculates the solvent accessibility of the active molecules. The Calculate Accessible Area dialog is used to specify calculation variables.
<i>Contact Area...</i>	This option calculates the contact area between two sets of atoms. The Calculate Contact Area dialog is used to specify the variables for calculating the contact area. The SelectSets palette is also activated to enable you to select two sets of atoms. When you enter this palette the Select Set 1 tool is active on the Selection Utilities palette and you should select the first set of atoms using the tools on the Select Sets palette. Then pick Select Set 2 and select the second set of atoms. Pick the Finish or Quit tools from the Select Sets palette to return to the Accessibility palette.
<i>Color by Atom</i>	This tool toggles coloring atoms according to their calculated accessibility.
<i>Color by Side Chain</i>	This tool toggles coloring side chains according to their calculated accessibility.
<i>Color by Fraction Accessible</i>	This tool toggles coloring residues according to their calculated accessibility as a fraction of their maximum possible accessible area.
<i>Color by Exception</i>	This tool toggles coloring residues that are in an unsuitable environment that is hydrophobic residues that have a high accessibility or hydrophilic residues that have a low accessibility. All other residues are colored gray.
<i>Color Options...</i>	This option displays Change Coloring Ranges dialog box. For a selected coloring mode you can change the colors used or the minimum accessibility cutoffs which define the color ranges.
<i>List to Textport</i>	This tool lists to the textport numerical information for either all atoms or all residues; those atoms or residues with zero accessibility are excluded from the list.
<i>Write to File</i>	This tool writes the accessibility numerical information to a file with the filename <i>MSFname_atom_access.out</i> or <i>MSFname_res_access.out</i> .
<i>Atom Accessibility</i>	This selects atom data for output. The format of the atom accessibility results is: atom number, atom name, segment ID and residue code, and accessibility of atom in square angstroms.

Residue Accessibility

This selects residue accessibility for output. The format for these results is:

- Residue number
- Segment ID and residue number
- Three-letter residue code
- Accessibility of the whole residue in angstroms squared
- Accessibility of the side chain in angstroms squared
- Fractional accessibility of the whole residue/side chain
- Indicator to show if the residue is in an exceptional environment

Excluded Zero Accessibility

This excludes from the output atoms and residues with zero accessibility.

Save to MSF

This tool saves all modifications and calculations to the active MSF as extra information with the label ACCESS and a title which you will be prompted to provide. Note that it is possible to save multiple sets of ACCESS data. The data saved with the title 'Default calculated accessibility' will be used to color the molecule whenever you use the Solvent Accessibility coloring mode. It will also be used, if it exists, in some Protein Health and Profile Analysis calculations so you should only save accessibility data, calculated for the whole molecule, with this title.

Reread MSF

This tool reads in extra information data labelled ACCESS from the MSF. If there is more than one set of data then you can select the set from the list of titles.

Finish

If you have done a calculation without saving the data to the MSF then you will be prompted to save it.

11.. Calculate Accessibility

12. Display Contact Maps

Overview

The most common use of contact maps is to show the inter-residue $C\alpha$ - $C\alpha$ distances. Similar properties that are a function of two residues, such as inter-residue van der Waals energy or number of inter-residue hydrogen bonds, can also be plotted in similar fashion. The properties currently displayed as contact maps in this utility are: $C\alpha$ - $C\alpha$ distances; $C\beta$ - $C\beta$ and side chain contact distances; van der Waals interaction energy; electrostatic interaction energy; total interaction energy; hydrogen bonds; and residue type interactions.

This chapter describes:

- Calculating Contact Maps
- The Plotting Method
- Distance Mapping
- Difference Mapping
- Energy Mapping
- Qualitative Mapping
- Tools and Options

For more information see:

Protein User's Reference

- Calculate Secondary Structure
- Protein Health

Calculating Contact Maps

Contact Maps in the Protein Design application uses basic x, y plots to analyze the relationship between pairs of residues. Three major categories of contact maps can be calculated: inter-residue distance, inter-residue energy, and residue-residue interaction type. In Protein Design, the contact maps for two proteins can be displayed side by side for easy comparison.

Plotting Method

The basic form of the plot for one protein of n residues consists of the x axis, representing residues 1 to n , and the y axis, representing residues n to 1. Therefore, the position (i, j) is color coded according to the value of the property between residues i and j . Since the properties are commutative, the value for (i, j) is the same as for (j, i) , then half of the plot is redundant. This allows the top-right of the plot to be used to display similar data for another protein.

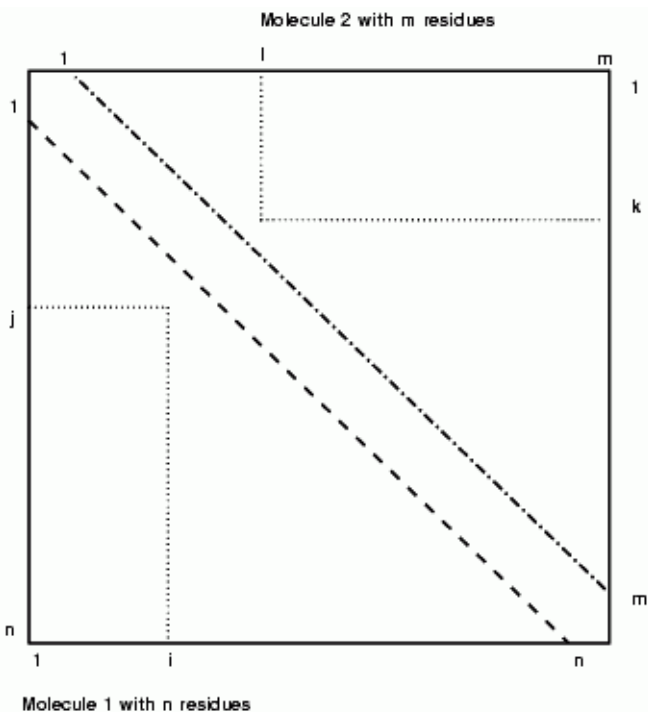
Molecule 1	1 i j n
Molecule 2	1 k l m

Sequence Table

The relationship between the sequence alignment and the contact map is shown in the Figures. In the sequence alignment two residues in molecule 1, i and j , are aligned with two residues in molecule 2, k and l . On the contact map, molecule 1 is plotted on the bottom left of the plot and molecule 2 on the top right. The position (i, j) on the bottom left of the plot shows the interaction between residues i and j in molecule 1 and, similarly, the position (k, l) on the top right of the plot shows the interaction between residues k and l . The two positions could be transformed one onto the other by a reflection in the bottom-right to top-left diagonal.

If the $C\alpha$ - $C\alpha$ distances of two homologous proteins are displayed side by side, the plot would have a rough mirror symmetry down the diagonal. However, if insertions or deletions are not taken into account in drawing the plot, the two halves of the plot would go out of step resulting in loss of symmetry. Therefore, to make comparison of two side by side plots easier, the axes of each plot incorporate any gaps in the sequence alignment. This

means that the plot may have black bands where there are gaps in the protein alignment.



Contact Map Diagram

This diagram shows the relationship between the contacts between residues i and j of molecule 1 and the residues k and l of molecule 2. The residues l and k are aligned and j , and i are similarly aligned.

A point on the plot can be picked by double clicking it. The actual residues and contact from that point are reported in the textport. If there are two contact maps on the plot, the same information is reported for the equivalent position on the other map.

Secondary Structure Elements

These are represented on the contact maps by rectangular boxes that enclose the inter-residue contacts between residues in an alpha helix or beta strand. The sides of the boxes are colored according to the type of secondary structure. Any contacts between residues within the same secondary structure element are close to the diagonal and enclosed in a three-sided box.

Molecule Display

The contacts that are shown on a map can also be displayed on the molecule as dashed lines of the same color as the point on the plot. They are located between either the C α or the C β atoms of the pair of residues. The amount of information can be undecipherable when displayed on a molecule. It is recommended that the option to select a limited set of residues be used for displaying contact information. The display can also be simplified by having a reduced representation of the protein, such as a C α trace, and toggling off the visibility of any irrelevant molecules.

Difference Contact Maps

Protein sequences should be correctly aligned before using the **Difference Contact Map** option.

The difference map is calculated by subtracting the contact map for the second molecule (upper right) from the contact map for the first molecule (lower left) with the result being displayed in the bottom left of the plot. The contact map for one of the molecules is shown in the top right of the plot. The exact interpretation of the difference plot depends on which property is being plotted.

The differences maps are colored for increasing magnitude of positive difference using pink (color 14) or red (color 3) and for increasing magnitude of negative difference using pale blue (color 12) and deep blue (color 2).

Distance Contact Maps

Three types of distances are mapped: C α -C α distance, C β -C β distance, and side chain contact distance. These maps are, by default, colored for increasingly close contacts, using white (color 5), pale yellow (color 6), deep yellow (color 4), and red (color 3).

The first two types of maps, C α -C α and C β -C β , are, by default, scaled to show fairly long range contacts, with all inter-residue contacts less than 16 Å shown. The maps show the overall folding of the structure, and usually have strong diagonal bands that correspond to two close strands in the structure. In this context, a strand might be secondary structure or extended coil. A band in the direction of the leading diagonal, bottom left to top right, correspond to anti-parallel strands, and bands in the direction of the other diagonal correspond to parallel strands.

A difference C α -C α distance map can show where there have been gross relative displacement of two strands. Where the contact distances are large

(the residues are far apart), are of less interest and can confuse the plot, they can be excluded from the display.

The side chain contact distance map shows the distance between the two closest atoms in the two residues side chains. This map has the same default colors as the $C\alpha$ - $C\alpha$ and $C\beta$ - $C\beta$ plot, but the default scaling is to show short range contacts less than 5 Å. This plot is useful for identifying interacting side chains.

By default, distance difference maps use the absolute difference coloring regime. It is also meaningful to color by the fractional difference, for example, the difference as a fraction of the smaller magnitude of the contact distance for the two molecules. This color scheme is accessed using the Contact Map Options dialog box from the **Display Contact Map** palette.

Energy Contact Maps

Three types of energy maps are available in this utility: van der Waals interaction; electrostatic interaction; total interaction energy. These energy contact maps will identify which residues interact and are colored for increasingly favorable, negative, interaction energies, pale blue (color 12) and deep blue (color 2). For increasingly unfavorable, positive, interaction energies, these maps use pink (color 14) and red (color 3).

The algorithm for the van der Waals energy is a simple 6–12 potential. The atomic radii are taken from the \$HYD_LIB/param.par file and depend on the selected atoms being correctly typed. The electrostatic calculation is a simple $q(i) \cdot q(j) / r^2$, with the atomic charges being those assigned to the atom and listed with the atomic information. The total interaction energy is the sum of the van der Waals and the electrostatic interaction energy. The energy calculations can optionally be restricted to the interaction between side chains and, by default, no energy is calculated for residues with the closest atom pair greater than 6 Å apart.

The energy difference maps are colored pink or red for increasing positive difference and pale blue and deep blue for increasing negative difference. By default, energy difference maps use the absolute difference coloring regime. As with distance maps (see “Distance Contact Maps” on page 88), the fractional difference coloring scheme can be used.

Interaction-Type Contact Maps

These two types of maps hydrogen bonds and residue types, are not of quantitative parameters but indicate types of interaction.

Hydrogen Bonds

Hydrogen bond interactions maps indicate whether H-bond interactions are between mainchain atoms, sidechain atoms, or (if multiple H-bonds exist) within a single atom. The classes of H-bonds and their default colors are:

- Mainchain–mainchain 10 (salmon)
- Sidechain–sidechain 12 (pale blue)
- Mainchain–sidechain 13 (rust)
- Multiple H bonds 14 (pale pink)

Residue Type

Residue type interaction indicates where residues have sidechains with atoms less than 6 Å apart. The indication of the type of the pair of residues is:

- Hydrophobic–hydrophobic 6 (pale yellow)
- Hydrophobic–hydrophilic 3 (red)
- Acidic–basic 12 (pale blue)
- Hydrophilic–hydrophilic 9 (light gray)

Tools and Options

The following section lists the tools and options found on the Display Contact Maps utility. It is possible to define a range in the sequence table using the **Select Active Range** tool from the **Protein Utilities** menu. When this tool is used the calculation is limited to the residues in the active range. This makes the axes range of the contact map limited to those residues.

If you want to compare two structures in side-by-side plots then their sequences must be aligned using the Align and Superpose utility.

Calculate Contacts...

This option displays the Contact Map dialog box, that enables you to select the type of contact map. This dialog box contains radio buttons for the different options, plus a toggle for calculating difference contacts.

Show Contacts Molecule 1

This option toggles the display of contacts on the first molecule in the viewing area. This corresponds to the contact information in the bottom left of the contact map. By default, this is the first molecule active on the Molecule Management Table. If more than one contact map has been calculated, the most recent one is displayed.

Show Contacts Molecule 2

This option toggles the display of contacts on the second molecule in the viewing area. This corresponds to the contact information in the upper right of the contact map. By default, this is the second molecule active on the Molecule Management Table. If more than one contact map has been calculated, the most recent one is displayed.

Change Displayed Contacts

This option displays the Change Displayed Map dialog box. When two or more contact maps have been calculated, the contacts displayed on the molecule are taken from the most recently calculated map. This dialog box enables you to select an earlier contact map.

Select one set of residues

This tool limits the interactions between a selected set of residues shown on the contact map or on the molecule display. The residue set must be selected before drawing and calculating a contact map. If the selection is changed while the contacts are displayed on the molecule, then the display is updated to reflect the new selection.

On the **Select Atoms** palette, one set of residues can be specified for calculating a contact map. Only contacts between the selected residues are displayed, but the axes of the contact map still have all the residues.

Only with themselves

This calculates and displays contacts between only the selected residues.

With non-selected residues

This calculates and displays contacts between the selected residues and the non-selected residues in the same molecule.

With all of protein

This calculates and displays contacts between the selected residues and the rest of the protein including the selected residues.

Select two sets of residues

This tool limits the interactions between selected sets of residues shown on the contact map and on the molecule display. The residue sets must be selected before drawing and calculating a contact map. If the selection is changed while the contacts are displayed on the molecule, then the display is updated to reflect the new selection.

On the **Select Atoms** palette, two sets of residues can be specified for calculating a contact map. Only contacts between the selected residues are displayed, but the axes of the contact map will still have all the residues.

Map Colors

This option displays the Contact Map Color Ranges dialog box from which the coloring schemes can be changed for the seven coloring regimes.* Each regime contains a set of ranges that can be edited from Define Coloring Ranges dialog box. For the first five coloring regimes, all distance or energy contacts with values within a given range are colored appropriately for that range.

*The seven types are: inter-C α and inter-C β distances; side chain closest contact distances; inter-residue energies; absolute differences; fractional differences; residue type difference; and hydrogen bonds.

12.. Display Contact Maps

The Define Coloring Range dialog box shows the maximum for each range and color.* The user can change the ranges or colors and the number of color bins.

Options

This option displays the Contact Map dialog box from which the default setting can be changed for various contact maps.

Distance and Energy Difference Maps

This determines if distance or energy difference maps can be colored according to either the absolute value of the difference or the difference as a fraction of whichever is the smaller in magnitude of the values for the two molecules. The default is **Absolute Differences**.

Show Absolute Values for which molecule

This determines if the first or second molecule is used to plot the differences between the two molecules on the bottom left of the contact map, and the absolute values for one of the molecules (by default the second molecule) are plotted top right. The default is **First Molecule**.

Use Distance cutoff in distance difference map

The default distance difference cut-off is 6.0 Å. When you look at distance difference maps, the difference in contact distance between residues that are far part is often not of interest and confuses the plot. With this option you can exclude distance differences where the separation distance on both molecules are greater than a given value.

Distance cutoff in energy calculation

This option sets the energy calculation for side chain only. In addition its use will speed up the energy contact map calculation.† The default is 6.0 Å.

Show contacts for core residues only

This option displays energy or distance contacts for core residues only, such as a fractional solvent accessibility from the default value of 0.5. The default is off.

List contacts to file

This option lists the calculated contacts to a file with name MOLECULE_contact_(CA/CB/energy/Hbond).out. This is done as the contact map is calculated. The default is off.

Show secondary structure on contact map

This option highlights in rectangular boxes the areas of the contact map showing interactions between residues in a secondary structure element, such as a alpha helix or beta strand. The default is on.

Finish

This tool returns you to the **Protein Design** palette.

*The range minimum either minus infinity for the first range or the maximum of the proceeding range. For those regimes that require the coloring of points with large positive values, the range maximum is specified as 9999.0, but interpreted as positive infinity.

†The calculation for pairs of residues is not done and atoms closer than the cutoff value are not used.

13. Analyze Domain Structure

Overview

Protein domains can be characterized, theoretically and experimentally, in several ways: by protein coordinates, by relative motions between domains, by the stability and folding of independent domains, or by different genetic origins and functions.

There have been several definitions of domains based on atomic coordinates. Domains can be defined in terms of:

- Using inter-C α distances
- Deriving a single cutting plane
- Minimizing the surface area of each domain
- Grouping of structural elements

The Protein Design application uses geometric relations between secondary structure elements to automatically identify domains, and provides tools that allow you to define and edit the domains.

This chapter describes:

- Analyzing Domain Structures
- Tools and Options

References

G. M. Crippen, *J. Mol. Biol.* **126** 315-332 (1978).

G. D. Rose, *J. Mol. Biol.* **134** 447-470 (1979).

S. J. Wodak & J. Janin, *Proc. Natl. Acad. Sci USA* **77** 1736-1740 (1980).

Analyzing Domain Structures

QUANTA describes a domain in terms of the secondary structure elements, rather than individual residues. A domain is defined as a group of close secondary structure elements and the loop regions are considered to be in the same domain as the secondary structure elements that they connect. The distance between two secondary structure elements is defined as the average distance between all pairwise combinations of C α atoms in the two elements. If this average distance is less than a given cutoff distance the

13.. Analyze Domain Structure

elements are considered to be in the same domain. The number of domains that the structure will subdivide into is dependent on the cutoff distance. For example, if the cutoff distance is decreased then fewer pairs of elements will qualify as being in the same domain and the protein will divide into more, smaller domains.

A simple clustering algorithm is used to analyze the distances between secondary structure elements and this generates a dendrogram. A dendrogram is a “family tree” of the secondary structure elements in which the pairs of elements which are closest in space are shown as most closely related in the tree.

Often the difficult step in domain analysis is deciding the appropriate cutoff distance for the inter-element average distance. The automatic algorithm will use a fixed value and report the number of domains which this will generate when you use the **Number of Domains** tool. You can alter the number of domains that are generated.

The Clustering Algorithm

This clustering algorithm finds the pair of closest secondary structure elements and joins them into one cluster. Then it repeatedly finds the closest pair of either individual secondary structure elements or clusters, which represent two or more elements. This continues until all the elements have been drawn together into one single cluster.

Clustering algorithms differ in how the distance from a cluster is calculated and how it is scaled compared with a distance from a single element. QUANTA's algorithm uses the distance from a cluster as an average of the distances from all the elements in the cluster. Therefore, the distance between two clusters is the average of all the distances between all the elements in one cluster and all the elements in the other cluster.

Associated with each cluster is a score that is the average of the distances between all the pairs of elements in the cluster. The result of the clustering is displayed as a dendrogram with the secondary structure elements listed down the screen. Elements or clusters that have been paired into a cluster are connected by a vertical line whose x-axis position is proportional to the cluster score.

Loop Regions

Residues in loop regions between secondary structure elements are assigned to domains using the following criteria:

1. Residues in loop regions between two secondary structure elements in the same domain are assigned to that domain.
2. For loop regions between secondary structure elements in different domains, a domain boundary is defined between two consecutive residues in the loop. The boundary is determined so as to minimize the sum of the distances from loop C α atoms to the nearest secondary structure element in the same domain. All residues in the sequence before that boundary are assigned to the proceeding domain and residues in the sequence after the boundary are assigned to the following domain.
3. N-terminal residues that are not in secondary structure elements are assigned to the next domain along the protein sequence. C-terminal residues that are not in secondary structure elements are assigned to the previous domain along the protein sequence.

Tools and Options

The overall structure of a protein can be better seen if you have only the C α atom trace displayed and colored according to secondary structure. The secondary structure elements can be highlighted by the **Secondary Structure** tool on the **Protein Utilities** palette. This shows a single vector for each element.

When this utility is used, only one molecule is active at a time. If more than one molecule is active when entering the utility, only the first remains active. If there is a domain definition saved to an MSF of a molecule, it is retrieved and used, otherwise a molecule is initially colored as a single domain.

All displayed molecules are colored to show their domain structure. For example, the first domain is color 1 (green) and the second domain is color 2 (blue). A legend on the bottom-right of the screen gives the molecule name and domain number in the appropriate color.

Three of the tools on the palette — **Number of Domains**, **One More Domain** and **One Less Domain** — automatically analyze the protein into some given number of domains. When these tools are selected, the molecule and sequence viewer are recolored to show the domain assignment of each residue.

There is a set of tools for manual assignment of secondary structure elements or individual residues to domains. If these are used, the molecule and sequence viewer coloring are updated appropriately, but the dendrogram coloring is not changed. If any of the automatic assignment tools are used after the manual tools, then the manual changes are overwritten.

13.. Analyze Domain Structure

<i>Display Cluster</i>	This tool toggles the display of a dendrogram.
<i>Number of Domains</i>	This tool displays the Enter Number of Domains dialog box from which to select the number of domains to be assigned to the protein. The maximum number of domains is equal to the number of secondary structure elements. The initial value in the dialog box is the automatic algorithm's best estimate of the number of domains using a fixed inter-element cutoff distance.
<i>One More Domain</i>	This option increases the number of domains by one. It is grayed out when the Number of Domains option has not been previously selected. When the maximum value has been reached no more domains are added.
<i>One Less Domain</i>	This option decreases the number of domains by one. It is grayed out when the Number of Domains option has not been previously selected. When one has been reached, no more domains are subtracted.
<i>Reassign Residue Range</i>	This tool displays the Pick Range palette from which to select a range of residues either off the sequence table or active structure. Once the range is selected, you are prompted to select a domain from a multiple choice list. The selected residue range will be assigned to the selected domain.
<i>Reassign Element</i>	This tool prompts you to select a domain from a multiple choice list, and to select an atom in the element that is reassigned to the selected domain.
<i>Create Domain</i>	This tool displays the Pick Range palette from which to select a residue range. Once the range has been selected, it is assigned to a new domain and given the next unused number.
<i>Merge Domains</i>	This tool displays the Pick Range palette from which to pick two residues, one in each of two domains that are to be merged.
<i>Undo Domain Edit</i>	This tool reverts the latest edit done on the domains.
<i>List Domains</i>	This tool lists to the textport the identity and residue range of the domains in the active molecule. The format is: Domain identifier — first residue in range — last residue in range
<i>Write Domains to File</i>	This tool writes domains to the file with filename MOLECULE_ domain.out. The format of the file is: Domain identifier — first residue in range — last residue in range
<i>Write Geometry to File</i>	This tool writes inter-secondary structure geometry to the file with filename MOLECULE_geometry.out. Listed for each pair of secondary structure elements is the structure type (such as B= beta strand or H= alpha helix), the ID of the first and last residue, the minimum and average distances between them, and the angle between them.
<i>Options</i>	This tool displays a dialog box to change the setting of the Cutoff Difference in Average Distance . The default is set to 2.5 Å.
<i>Save to MSF...</i>	This tool saves domains as extra information to the MSF.

Reread MSF...

This tool displays a dialog box that has extra information titles from which to read the domain structure.

Finish

This tool removes the **Domain Analysis** palette and returns the **Protein Design** palette. If the domain structure has been changed, a dialog box for each structure is displayed offering the option of saving domain structure information to the MSF.

13.. Analyze Domain Structure

14. Profile Analysis

Overview

Profile Analysis can either be activated from the **Protein Design** palette or the QUANTA **Applications** menu. When activated from the **Application** menu the **Protein Utilities** menu is also displayed. Profile Analysis follows the method of Bowie, Luthy and Eisenberg in analyzing protein structures into 1D profiles which can be assessed against protein sequences to quantify the quality of a structural model.

This chapter describes:

- Analyzing Protein Profiles
- Tools and Options

References

U. Bowie, R. Luthy & D. Eisenberg “A method to identify protein sequences that fold into a known three dimensional structure” *Science* **253** 164-170 (1991).

R. Luthy, J. U. Bowie & D. Eisenberg “Assessment of protein models with 3D profiles” *Nature* **356**, 83-85 (1992).

Analyzing Protein Profiles

Generating a 1D Profile from a 3D Structure

Using this method, the environment of each residue in the protein structure is analyzed in terms of its secondary structure and environment, then a profile sequence is generated in which each residue is assigned to one of 18 environment classes.

The definition of residue environment is a function of two parameters: its side chain buried area and the polar environment of the side chain. The buried area of the side chain is defined as the difference between the solvent accessible area of the side chain and the maximum possible solvent accessible area. The maximum solvent accessible area is defined as the solvent accessible area for the side chain in the tripeptide of GLY-X-GLY when it is in a fully extended conformation; in this situation there are no other residues to occlude the central X residue.

The polar environment of a side chain is the proportion of the side chain area which is covered by polar atoms which can be either solvent or polar atoms abutting onto the side chain surface.

Three categories of solvent accessibility are defined as: E (exposed), P (partially buried), and B (buried). Dependent on the fraction of the environment which is polar atoms the partially buried category is further subdivided into two categories and the buried category is sub-divided into three categories. These are designated P_1 , P_2 , B_1, B_2, B_3 where the higher subscript denotes a greater polar environment. Combining the three recognized secondary structure types and these six side chain environment categories gives 18 possible residue environment classes.

Comparing a Profile to a Sequence

A profile sequence is similar to a conventional sequence except that it lists residue environments rather than amino acids. From analysis of known structures it is possible to determine a quantitative score for the preference of each of the 20 amino acids for any of the 18 residue environments. With this means of scoring the suitability of an amino acid to a given residue environment, it is possible to do a conventional sequence alignment of an amino acid sequence to a profile sequence. An alignment score can then be calculated to give some measure of the suitability of that amino acid sequence to the profile sequence.

Plotting Profiles

Calculating a structure profile requires several fairly time consuming calculations of residue buried area and polar environment and once these are calculated they are usually saved to the MSF file as extra information and retrieved whenever the profile of that structure is required in future. The **Plot Structure Profile** tool will calculate the profile for a structure (or restore it from the MSF, if possible) and generate a graph for the structure's own sequence assessed against its own profile. This plot is an indication of the quality of the model with a score for each residue in the structure. It is conventional to integrate the residue scores using a window of the order of nine residues as this produces a plot which is easier to interpret.

Comparing Profiles to Other Sequences

To compare profiles with other sequences, you should use the **Select Sequence** tool to select one sequence. It is then possible to use **Plot Sequence Profile** to generate a graph showing the score of the sequence against all currently active structures with profiles. This tool will assume the current alignment between sequence and structure(s). It is possible to attempt to optimize the sequence—structure alignment using the **Align** and **Dot Plot** tools.

Tools and Options

Within Profile Analysis are tools to calculate the 3D profile for a selected structure. The calculation of buried areas and polar environments is slow. Therefore, once a profile analysis has been calculated, it is automatically saved to the MSF as extra information with the titles:

- Default residue buried area;
- Default residue polar environment;
- Default 3D profile environment.

Once a profile has been calculated the molecule is colored according to the residue environment class. The **Protein Utilities Legend** tool can be used to toggle the display of the color legend.

Plot Structure Profile

For all currently active structures the residue buried area, residue polar environment and secondary structure are calculated and the 1D profile is derived from these data. This information is saved as extra information in the MSF. If the information is already present in the MSF then this is used and it is not recalculated. The assessment of the structure against profile for each active molecule is plotted to the sequence viewer. The window parameter used for the plot is controlled by the **Profile Options** tool.

Select Sequence

You are prompted to select one sequence which may be an MSF or a sequence without an MSF. The selected sequence will be assessed against active molecules with profiles by the **Plot Sequence Profile** and **Dot Plot** tools.

Plot Sequence Profile

For the current selected sequence and all active MSF structures the assessment of the sequence vs. the structure profile is plotted to the Sequence Viewer. If there is more than one active structure, then there will be more than one plot and these have the names of the structure in the graph legend area. The legend title includes the name of the sequence.

Dot Plot

Dot plots are explained in Chapter 5. The dot plot parameters of window length and color range can be changed by the **Options** tool. The dot plot shows the current sequence against one structure profile and indicates possible alignment of sequence and structure by the stronger diagonal lines. A dot plot of a structure profile against its own sequence for a “good” structure will show the quality of data that can reasonably be expected with this method.

Align

The current sequence is aligned against one active structure profile. The gap penalty used in this context should probably be small to correspond to the low scores that usually result from the scoring. The gap penalties can be changed by the **Options** tool.

14.. Profile Analysis

Undo Alignment

Remove any gaps in the alignment of all active sequences.

Recalculate Profile

By default, once a profile has been calculated and saved to MSF by the **Plot Structure Profile** tool it will be used in all future assessments and plots. This tool will enable recalculation of the structure profile which may be required if the structure has been changed.

Options

The adjustable parameters are:

- The profile plot window which determines the number of residues over which sequence versus profile plots are integrated.
- The alignment gap penalties.
- The window length and color range cutoffs used in dot plots.

Save to MSF

This tool saves the current calculated profile to the MSF. The standard MSF saving options are displayed.

Read from MSF

This tool rereads the last saved version of an MSF and makes it current.

15. Protein Information

Overview

This utility retrieves textual information on PDB files from the protein structure database by accessing the QUANTA file \$HYD_LIB/data-base.dat. This database file contains information on all the PDB files currently in the Brookhaven Protein Databank. It is the same data file used by the structural database utility.

An example of how this utility might be used would be to query for information on all the hemoglobins in the database. The query would return a list of all the hemoglobin PDB files, a short textual description of each, and data, such as the number of residues in the PDB file.

This chapter describes:

- Retrieving Protein Information
- Tools and Options
- Running a Protein Information Query

References

IUPAC-IUB Commission on Biochemical Nomenclature.

Biochemistry **9** 3471-3479 (1970).

C.M. Wilmott and J.M. Thornton *J. Mol. Biol.* **203** 222-223 (1988).

W. Kabsch and C. Sanders *Biopolymers* **22** 2577-2647 (1983).

Retrieving Protein Information

This utility retrieves textual information for each protein structure without reading in the protein structure. It is activated from the **Protein Design** menu and displays the Specify Proteins dialog box. The dialog box contains five options, several with preset defaults.

Once the PDB textual information has been retrieved, it is listed to the textport. This information includes:

- Full protein name
- Structural family
- Ligands

- Number of segments
- Residues
- Solvent molecules

Tools

Search for keyword or PDB

This option displays the Specify Protein dialog box. Using different variables within the option fields, it is possible to retrieve either general or specific information.

Each of these options is described below.

This tool allows you to enter either a keyword or the PDB filename for the protein structure of interest. More than one keyword or PDB name can be entered by clicking the **OK** button. The text already entered is saved, and the entry field cleared so additional text can be entered.

If more than one keyword or PDB name is entered, these are considered to be connected by a logical OR. If two keywords are entered, then information will be retrieved for any protein that has either keyword_1 or keyword_2.

Maximum crystal resolution

This tool enables you to limit the search to structures whose crystal structure determination had a resolution less than a given value. The lower the resolution, the better the structure. For example, a resolution less than 2.0 Å is good. However, less than 3.0 Å may be acceptable for determining the main chain conformation and side chain position, but some parts of the structure may not be resolved as well as others.

Position in database between structure number

The search normally looks through every protein in the database, but this tool enables you to limit the range of proteins searched. This option is only useful if the your database is set up with a known selection of proteins in a particular position in the database.

Output log file name

This tool enables you to specify an output log filename or the default name, info.log. Once the **Search** button is clicked, a command file for the search program is written and the search runs automatically. The results are written to the selected log file and also displayed in the textport. The default the log file is automatically overwritten each time it is used.

Running a Protein Information Query

The following exercise demonstrates how to use the Protein Information utility and shows an example of typical information results.

1. From the **Protein Design** menu select the **Protein Information...** option.

The Specify Proteins dialog box is displayed.

2. Enter the following variables:

Search for keyword: **pepsin**

Maximum crystal resolution: **5**

Output log file name: **pepsin.log**

3. Click the **Search** button. The search is run and the results are displayed in the textport. Press <Enter> to continue scrolling through the information in the textport. To quit, press <q>, and then <Enter>. The information is automatically stored in the file pepsin.log for use later.

```

#
# PDB file name: 2ER7
#
# Protein name: ENDOTHIA ASPARTIC PROTEINASE (ENDOTHIAPEPSIN) (E.C.3.4.23.6)
# COMPLEX WITH H-261
#
# Protein family: HYDROLASE (ACID PROTEINASE)
#
# Crystallographic resolution: 1.6A
# Number of molecules in assymetric unit: 0
# Total number of segments: 2   Total number of residues: 336
#
#
# Segment id   No res   First res   Last res
#   E         330     E-2        E326
#   I          6      I2         I8
#
# Ligands:
# id name description
# BOC I 1 7 TERT-BUTYLOXYCARBONYL ANION
# LOV I 6 15 LEU-PSI(CHOH-CH2)-LEU
# SO4 900 5 SULFATE
# SO4 910 5 SULFATE
# SO4 920 5 SULFATE
#
MSF does not exist for protein: 2ER7

```


16. Sequence Database

Overview

This utility provides a group of options to search the Protein Data Bank for sequences that closely match a specific sequence. QUANTA uses the FASTA sequence search algorithm.

This chapter describes:

- Reading a Protein Sequence

For more information see:

Protein User's Reference

- Align and Superpose
- Appendices: Reading Sequence Formats
- Read Sequence/Alignment File

References

D. J. Lipman and W. R. Pearson, "Rapid and Sensitive Protein Similarity Searches", *Science*, **227**, 1435 (1985).

M. O. Dayhoff, *Atlas of Protein Sequence and Structure*, (National Bio-medical Research Foundation, Silver Spring, Md., 1978), volume 5, supplement 3

FASTA Sequence Searching

The Sequence Database application in QUANTA provides a sequence search option, searching the Protein Data Bank for sequences that closely match a specified sequence. The FASTA sequence search algorithm* is used to search protein sequences stored in the file \$HYD_LIB:pdbsequence.lib.

When FASTA initially scans the protein sequence library, the target sequence is compared to each library sequence. The best region of homology without gaps is found and an initial score is calculated. If the initial score is greater than the cutoff score, the library sequence is stored for later consideration.

*D. J. Lipman and W. R. Pearson, *Rapid and Sensitive Protein Similarity Searches*, *Science*, **227**, 1435 (1985).

16.. Sequence Database

The cutoff score used in the FASTA search is automatically calculated and is dependent on the query sequence. If the reference sequence is short, say less than 28 residues, then it may be necessary to specify a cutoff score since the cutoff generated automatically is usually too severe. When searching for short sequences, use a fairly high cutoff score initially and if this fails to find any matches, use a lower cutoff score.

Database search results are automatically listed to the QUANTA textport and written to the output file .log. The results include:

- A histogram indicating the number of database sequences found against their initial score
- The names of protein sequences that have the highest scores
- Protein sequences showing the alignment between the query sequence and the retrieved sequence

Tools

This utility displays the Sequence Database dialog box. All the options activate the File Librarian.

Enter search sequence

You are prompted to give a file name for a sequence file and then prompted to enter the sequence as single letter or three letter amino acid code.

Enter a blank line to terminate sequence input. The sequence is written to a FASTA sequence file with file extension .fta. Alternative means to generate the same file are using the **Create Sequence** option in the Protein Editor or using the **Write Sequence File** option of the Sequence utility on the **Files** pulldown. The latter option will write out any currently selected sequence, be it sequence-only data or from an MSF, to a FASTA format file.

Run sequence search

You are prompted to select a sequence file and the search job is run. The results are displayed in the textport and also saved to a log file.

Read sequence search log file

This option displays the File Librarian to select a .log file to read. Once the log file is selected, the results are displayed to the textport. To browse the file, use the <Enter> key to display the information to the textport and the sidebar to move up and down.

Change sequence database file

This option displays the File Librarian to select an alternate sequence database file to use. The default database file used is pdbsequence.lib.

17. Structural Database

Overview

This utility is a graphical interface to a protein database search program. The database query is specified by entering the data in dialog boxes. The protein fragments which satisfy the search query are read into QUANTA and displayed.

The database information is derived for all the known protein structures from the Brookhaven Protein Databank and stored in the file \$HYD_LIB/database.dat. The stored structural information is primary residue-based and includes some information on atoms and secondary structure. Searches to this database can address many common modeling problems. However, facilities for statistical analysis are not provided at this time.

structural
database

This chapter describes:

- Searching the Structural Database
- Defining a Structure Database Query
- Tools and Options Definitions

For more information see:

Protein User's Reference

- Customizing the Databases
- Database Troubleshooting
- Running the Search Program Stand Alone
- Side Chain Torsion Angles and Centers
- Wildcard Residue Type File

References

IUPAC-IUB Commission on Biochemical Nomenclature.

Biochemistry, **9**, 3471-3479 (1970).

C.M. Wilmott and J.M. Thornton *J. Mol. Biol.*, **203**, 222-223 (1988).

W. Kabsch and C. Sanders *Biopolymers*, **22**, 2577-2647 (1983).

Searching the Structure Database

The protein structural information has been extracted from the current Brookhaven Protein Databank and stored in a single file usually called \$HYD_LIB/database.dat.

For each protein the stored information consists of:

- Name of the Brookhaven PDB file
- Text description of protein name, family, crystallographic form, and refinement method
- Number of segments, residues, atoms, and solvent atoms
- Ligands residue name, description and number of atoms

For each residue the stored information consists of:

- Segment name
- Residue name and type
- Secondary structure type^{*}
- C α atom coordinates
- Sidechain “center” coordinate
- Pseudotorsion between four successive C α atoms in a chain
- Mainchain phi and psi angles
- Sidechain torsion angles

Sidechain torsion angles and centers are described further in Appendix H.

You can append additional proteins to this file or create your own separate database file as described in Appendix C.

A program external to QUANTA, called SEARCH, is used to search the structural database file. This program can be run stand-alone using a control file created within QUANTA or edited manually. The format of the control file is described in Appendix G.

Defining A Structure Database Query

Four types of information can be used to define a structure database query:

^{*} Kabsch and Sanders method is used to determine secondary structure type.

Proteins to search

The search can be limited to those proteins with a resolution better than some specified value or using a keyword such as “hemoglobin” in their description. However, by default all proteins are searched.

Templates

A template is a fragment of consecutive residues. Each residue may be defined by residue type (or some wildcard group of residue types such as “hydrophobic”, by secondary structure type, by main chain or side chain torsions or by intra-template C α -C α distance. Each residue in the template may be defined as much or as little as required.

Constraints

A constraint is a limiting relationship between two templates. It specifies that the two templates must be a certain distance apart or a certain number of residues apart in the sequence. For example:

- A limit of maximum and minimum allowed distance between C α atoms and/or sidechain “center”
- A limit on the maximum and minimum number of residues in sequence between residues in different templates

If two or more templates are defined, they must be linked by constraints, otherwise they are completely independent and should be treated as separate database queries.

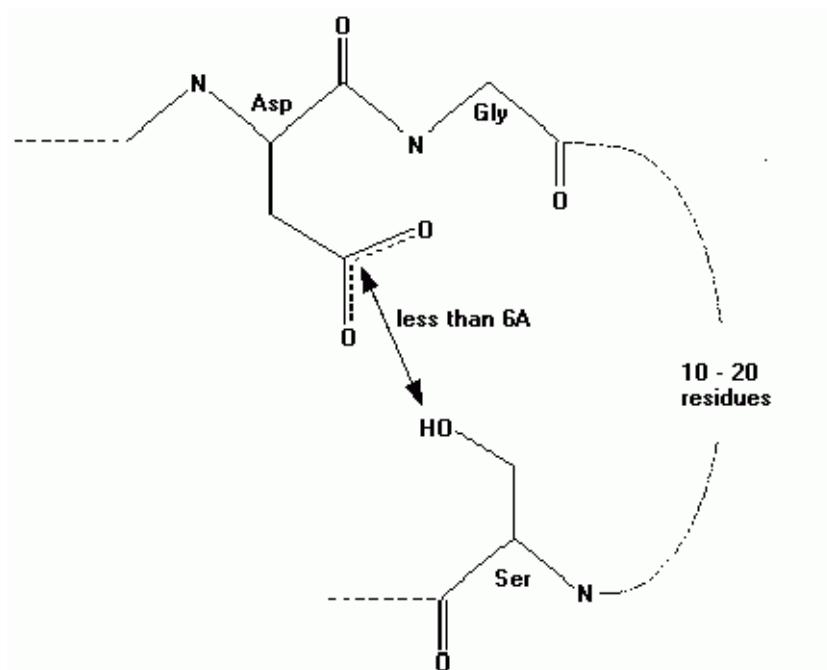
For example, to find all instances of the two residues asp and gly where the asp residue is near a ser residue that is between 10 and 20 residues ahead of the asp residue in the sequence, then the form of the query is:

Template 1: A two residue fragment with sequence asp-gly

Template 2: A ser residue

Constraint 1: The side chain of the first residue in template 1 (asp) is between 3 Å and 6 Å away from the sidechain of the first residue in template 2 (ser).

Constraint 2: The first residue in template 2 (ser) is between 10 and 20 residues ahead of the first residue in template 1 (asp) in the protein sequence.



A database search usually finds a number of hit fragments that satisfy the query. These fragments, from different proteins, are usually superposed for easy comparison and then displayed within QUANTA.

Search parameters

The user can specify how tightly the search criteria must be adhered to by changing the search tolerance. For example, if you specify a required main chain or side chain torsion, then, by default, all structures within 30° of the required value will be retrieved. After you enter the required information using the **Define Search** tool, the control file for the search program is generated, and the search runs automatically.

The search program produces two output files: a log file that lists all hits, and a selection file. The selection file is in standard selection format, and is used by QUANTA to read in the required fragments taken from a library of MSF files. This library is a directory containing MSFs, and is pointed to by the environment variable \$MSF_LIB. Each file should correspond to an equivalently named Brookhaven PDB file.

While reading in the fragments, it is optional to read in additional neighboring residues in order to display the environment of the fragment of interest. Fragments from different molecules are usually superposed so they can be compared more readily. The fragments taken from multiple MSFs are writ-

Browsing Database Search Results

ten to one MSF (along with some additional information on the database search).

The database browse tools allow rapid display and comparison of the hit fragments from one database MSF (i.e. an MSF containing fragments of multiple proteins which were found in the database search). They are the same as the browse tools found in the Fragment Database palette from the Model Backbone utility.

If more than one database MSF is selected, you are prompted to choose one to browse. When the search results are read in the MSF created automatically becomes the browse MSF. The jobname of the database search of the current browse MSF is printed at the bottom right of the screen. Each group of residues that constitute one hit are in a single segment, and the segment names that identify the fragments are displayed on the right of the screen. The segments are named with a three-letter code for the PDB file that they came from, and a single-letter code to differentiate the hits from the same PDB file.

structural
database

Tools and Options

Define Query...

This tool opens a series of dialog boxes for conducting a search. The first dialog is the Define Protein Structure Database Query.

A query must have at least one fragment template defined. Each additional template should be related to the existing template by a constraint. Otherwise, the templates are effectively independent and should be searched for in separate jobs. Errors in defining a template or constraint can be corrected using the **List/delete** option. The proteins to search and search parameters need not be entered, as the defaults are usually satisfactory. After finishing the definition of the query pick the Search button to initiate the search.

Define template

This option open the Define Template dialog box.

Enter the number of residues being searched for and a name for the template. The template name is used later for reference. If side chain torsions or intra-template distance constraints are required, toggle on the appropriate buttons. This will activate the appropriate dialog boxes that define these parameters.

*Define template name_
search*

This dialog box enables you to define the template residues.

It has the appropriate number of lines for the number of residues, and columns for residue type, secondary structure type, and phi, psi, and C α pseudo-torsion angles. The residue type can be set to the wildcard ANY and boxes left blank are considered undefined.

Valid residue types are the 1 or 3 letter code for the amino acid or a wildcard. The default wildcards are defined in the file \$HYD_LIB/wildcard.dat, and can be listed to the textport by toggling the option in the Define Query dialog box. To specify the wildcard type, enter the wildcard code (a maximum of six-letters) in the residue type field. The allowed wildcards can be extended or changed by editing the wildcard.dat file.

Valid secondary structure types are:

- E— Extended chain
- H— Folded conformation (including the next 4 types)
- 3— 3-turn
- 4— 4-turn
- 5— 5-turn
- A— α -helix
- T— Turn (includes 3,4, and 5)

Each of these can be prefaced by NOT. The secondary structure types in the database are designated using an algorithm based on Kabsch and Sanders that is identical to the one used in QUANTA.

The valid torsion angles are in the range -180° to 180° . The torsion parameters are target values; a hit will be any residue that has torsions within the tolerance angle of that target value. The tolerance angles can be set within the **Search Parameters** options; the default tolerance is 30° .

Side Chain Torsions

If the sidechain torsions option has been selected this dialog box opens to define sidechain torsions.

As with the mainchain torsions, you enter a required target value, and all structures within the specified tolerance (default 30°) of that target value are retrieved.

Define INTRA-template constraint

If the **Intra-residue distance** constraint option has been selected, this dialog box opens to define intra-template $C\alpha$ - $C\alpha$ distance constraints.

This option is for constraints between residues in the same template. Choose the type of constraint, a target distance, and two residues are from the two scrolling lists. Note: This option is very different from the **Define** constraint between two templates option, which is available from the Define Protein Structure Database query dialog box.

The database contains only limited atom coordinate information: the $C\alpha$ atom coordinate for each residue, and a coordinate for the side chain center. The definition of the side chain center for each amino acid type is listed in Appendix, Side Chains and Torsion Angles. This limitation means that a distance constraint between two residues can be only one of three types: a

distance between two C α atoms, a distance between a C α atom and a side chain center, or a distance between two side chain centers. An appropriate constraint type is selected, then enter the required target distance, the default tolerance is 1.0 Å, which can be changed with the Search parameters option. Particularly when searching for distances between two side chain centers it may be necessary to search with generous tolerances on the distance criteria in order to find all required structures; structures that are not required can be rejected later after inspection.

If you click the **OK** button, the currently displayed definition is saved and the dialog box options are reset to the initial default values so that you can enter further constraints. Clicking the **Quit** and **Finish** buttons removes the dialog box. However, the current definition is saved only if you click the **Finish** button.

Define constraint between two templates

This option opens the Define Constraints Between Two Templates dialog box.

A constraint is a relationship between two templates. If you define more than one template, each additional template must be related to at least one of the existing templates by a constraint. Otherwise, the templates are independent and could be treated as separate queries. More than one constraint can be defined between a pair of templates. A constraint can be either a distance constraint, (e.g. the distance between a C α atom in one template and a C α atom in the other template), or a residue separation constraint which specifies that two templates are separated in the sequence by a given number of residues.

Using Distance Constraints

The database contains only limited atom coordinate information: the C α atom coordinate for each residue, and a coordinate for the side chain center. The definition of the sidechain center for each amino acid type is listed in Appendix H. This limitation means that a distance constraint between two residues can be only one of three types: a distance between two C α atoms, a distance between a C α atom and a side chain center, or a distance between two side chain centers. An appropriate constraint type is selected, then enter the required target distance, the default tolerance is 1.0 Å, which can be changed with the Search parameters option. Particularly when searching for distances between two side chain centers it may be necessary to search with generous tolerances on the distance criteria in order to find all required structures; structures that are not required can be rejected later after inspection.

Using Residue Separation Constraints

The obvious form of residue separation constraint is to specify a variable number of residues between two fixed patterns of residues. For example, to find a sequence pattern G-G-(2,5)X-G—two consecutive gly residues followed by between two and five residues before another gly residue—the first template is two gly residues and the second template is one gly residue.

The constraint is that the first residue of the first template and the residue of the second template are between four and seven residues apart.

It also is often necessary to set an exclusion range of residues. For example, in studying side chain-side chain interactions you may require that the interactions be between residues remote in sequence. Two templates can be defined each containing one residue of the amino acid type of interest, and set an exclusion range between the two templates. If this range is set between -5 and +5 residues, then the two residues must be at least five residues apart in the sequence.

An instance where it is necessary, but not obvious, to set a constraint is if two identical templates are defined. For example, in searching for two interacting histidine residues, you would define two templates; one for each histidine residue. In this case you also must specify an exclusion range between the two templates of at least -1 to 1 to ensure that any single histidine residue does not satisfy both templates.

In other instances, it may be desirable that one residue in a protein is simultaneously in two templates. Take, for example, search for a structure with two α -helices which are connected by a loop of between 6 and 12 non-helical residues. To do this you would define two templates and set a residue separation constraint between them. The first template is 12 residues long with the first six residues specified to be helix, and the second six specified as not helix. The second template is 12 residues long, with the first six residues not helix and the second six helix. The constraint is that the first residue of the first template, and the first residue of the second template are between 6 and 12 residues apart.

This could be illustrated by the two extreme solutions: where H = helix, and N = not-helix.

First residues in the two templates separated by 6 residues:

Template 1 H-H-H-H-H-H-N-N-N-N-N
Template 2: N-N-N-N-N-H-H-H-H-H

Fragment found: H-H-H-H-H-H-N-N-N-N-N-H-H-H-H-H

First residues in the two templates separated by 12 residues:

Template 1: H-H-H-H-H-H-N-N-N-N-N
Template 2: N-N-N-N-N-H-H-H-H-H

Fragment found: H-H-H-H-H-H-N-N-N-N-N-N-N-N-N-N-H-H-H-H-H

In the case which the first residues area separated by 6 residues, some residues in the hit structure are in both templates.

Specify proteins to search

This option opens the Specify Proteins to Search dialog box used to specify proteins for searching. It is the same dialog used for Protein Information.

By default, any query will search the entire database until it has found the required number of hits. It is possible to limit the search to particular proteins and it is often desirable to limit the search to structures of higher resolution.

Search Parameters

This options opens the Database Search Parameters dialog box.

In defining a search template, you enter target values for torsion angles and distances. Structures with geometry within some given tolerance of the target parameters are retrieved. The database search will end after finding some specified number of hits; the default is 50. The tolerance on the target parameters and maximum number of hits retrieved can be changed. The name of the protein database and wildcard files can also be changed. If you change these parameters, the new values remain operative through the current QUANTA session, and are saved in the protein constants file for use in subsequent sessions.

List secondary structure and amino acid wildcards

This option lists to the textport the one letter codes for secondary structure types and the amino acid wildcards currently defined in the wildcard.dat file.

List/delete existing template(s) and constraint(s)

This options opens the Delete Template or Constraint dialog box, which lists the current names of templates and constraints.

Click the **Quit** button to exit without any changes to delete a template, make a selection and click the **Delete** button. If a template is deleted the default is that any constraint involving that template is also deleted. This can be switched off by clicking the **Delete Constraints** toggle at the bottom of the dialog box.

Run Search...

The database search program, \$HYD_EXE/search, is a separate executable which is run from within QUANTA with this option. The results are read automatically back into QUANTA. Each database search is given a job name. This name is used as a root name for the files used to transfer data between the two programs. The files are of the format: job_name.ddb. This is the query file generated by QUANTA and read by the search program. The internal format of this file is described in the Appendix, Running the Search Program. The file can be edited, and the search program run outside QUANTA.

The search program outputs two files: *job_name.log* is a log file that lists all the hits and reports any problems; *job_name.sel*, which is in the format of a QUANTA selection file, specifies which fragments of selected proteins should be read into QUANTA for display. The log file can be reviewed from within QUANTA using the Read log file option.

If you have just entered a database query and clicked the **Search** action button, you are prompted for a jobname. The query is written to the file *job_name.ddb* and the database search initiated. If the **Run Search** option is

picked without entering a query, you are prompted to select an existing query file (.ddb). QUANTA waits while the search is run. It normally takes a couple of minutes to search the entire database with a simple structural query. Once the job is finished the results are read and displayed in the textport.

Read Search Log...

This tool opens the File Librarian to select a log file. This log file from the database search is listed to the textport. The file lists the proteins with structures that satisfy the query and the residues that correspond to the query templates. The log file also reports if the MSF file containing a hit fragment cannot be found. If you run a database query then the log file from that query is automatically listed; otherwise you are prompted to select a log file.

Read in database search results

In order to display the results of the database query, the fragments of proteins that satisfy the query are read into QUANTA. They are then written out to a single MSF with some additional information relating to the database search in the title records of the MSF. MSFs that have been created using this utility can be displayed, and can be reviewed using the database browse facility. They are handled appropriately in other applications of the Protein Design application where the coordinates or conformations of the fragments can be copied to modeled structures.

The selection file, job_name.sel, specifies which residues of selected proteins satisfy the database query. The structural information for these fragments is read from an MSF for the protein. If the database search has found hits in protein structure for which there is no MSF file, either in the directory \$MSF_LIB or in your working directory, then these hits are not included in the selection file but are reported in the log file. To display these hits an MSF will need to be created by reading in an appropriate PDB file. The MSF can be put in either the MSF library directory or in your own working directory. The database search can then be rerun so the selection file includes the relevant MSF.

If the **Read Hit Fragments** tool is selected without having just run a database query, you are prompted for the name of a selection, .sel, file to read. To run successfully the query file, job_name.ddb, must also be present.

Read Hit Fragments

This tool displays the Select Fragment to Display dialog box. It reads in and, by default, displays all the hit fragments referenced in the selection. It also enables a limited selection to be specified and read. The default setting only reads those residues corresponding to the residues of the query templates, but you can choose to read in neighboring residues.

Select hits to display

This option, by default, reads into the initial MSF all the hit fragments. Alternatively, a range can be selected with this option. For example, selecting the range 1 to 20 would read in the first 20 hits in the selection file. In

addition, specific hits can be selected from the scrolling list of all the protein MSF files in which hit fragments have been found.

Select extra neighboring residues to be displayed:

This option displays two choices. They are either to select, a zone of residues, such as selecting residues contiguous with a query template, or a sphere, such as all residues close to a residue in a query template. If either of these options is picked, the appropriate dialog box to enter the selection is displayed.

Select zone

This selects an extra zone of residues for each template. In selecting a zone, choose the number of residues before the first template residue, then number of residues after the last template residue. The default is zero residues, that selects no additional residues.

Select sphere

This selects a sphere of residues. In selecting a sphere, choose one of the template residues from the scrolling list as the center of the sphere, then enter a radius for the selection sphere. The selection uses the same criteria as NAYBR BYRES, so that all residues that have an atom within the cutoff distance of the center residue are selected. The selection is applied if you click **OK** or the **Quit** button. If you click **OK** the dialog box remains and another selection can be made. Click **OK**, for example, if you have two templates and want to select an extra sphere of residues about both templates.

Superpose hits

This superposes hits and makes it easier to compare the hit fragments. Once the hit fragments are read in and a new MSF created by default you are automatically prompted to select the atoms to superpose. However, this option can be toggled off.

The fragments read in are all written out to one MSF. Each group of residues that constitute one hit are in a single segment. The segments are named with a three-letter code for the PDB file that they came from, and a single-letter code to differentiate hits from the same PDB file.

Write to default MSF name

This tool opens the File Librarian after a search has been completed for writing the hit fragments to file. The hit fragments, by default, are written to an MSF called job_name.msf, but an alternative filename can be used.

Superpose Fragments...

This tool superposes fragments after you have read in hit fragments and created a new MSF. By default you are prompted to superpose the fragments. It is also possible to superpose the fragments later on using the **Superpose Fragments** tool.

The superposition is a conventional least squares superposition of equivalent atoms, as described in the Align and Superpose module. It depends on the context in which atoms are most appropriate to superpose. However, it is usually easier to interpret the display if you have a good superposition of one residue, by selecting just three or four atoms from that residue, rather than having a rough fit over several residues.

17.. Structural Database

The dialog box to select atoms to superpose has a scrolling list of template residues. Select a residue and a text input into which you should enter the names of the atoms to superpose. If you wish to select atoms from more than one residue, click the **OK** button to enter the current selection, and then enter your additional selection. Click the **Superpose** button to initiate the superposition.

Save to MSF

This is grayed until the fragments have been superposed. This tool saves the superpositioned fragment to the MSF. The standard saving dialog options are displayed.

Restore from MSF

This is grayed until the fragments have been superposed. This restores the previous coordinates of the fragments and the superposition is discarded.

Display All Fragments

This tool displays all hits superposed over the structure.

Display Next

This tool selects the next available fragment until it has reached the last one and then it is grayed out.

Display Previous

This tool selects the last fragment until it has reached the first one and then it is grayed out.

Select Display

This tool opens the Display Selected Fragments dialog box from which to select fragments to display.

Display Option

This tool opens the Fragment Display Mode dialog box from which to select the **Display** options.

18. Motif Database

Overview

This module provides tools that search a database for structures with similar folds to one active molecule. The reference motif, by default, is all the secondary structure elements of the active molecule. However, it is possible that elements can be set inactive and excluded from the reference motif.

This chapter describes:

- Searching the Motif Database
- Tools and Options Definitions
- Motif Database Log File

For more information see:

Protein User's Reference

- Superpose Motif
- Domain Analysis

References

“Use of techniques derived from graph theory to compare secondary structure motifs in proteins,” E. M. Mitchell, P. J. Artymiuk, D. W. Rice & P. Willett, *J. Mol. Biol.* **212** 151-166 (1989).

“Pharmacocophoric pattern matching in files of 3D chemical structures: comparison of geometric searching algorithms,” A. T. Brint & P. Willett, *J. Mol. Graph* **5** 49-560 (1987).

“Correction to Bierstone's Algorithm for Generating Cliques,” G. D. Mulligan & D. G. Corneil, *J. ACM* **19** 244-247 (1972).

“Fast Structure Alignment for Protein Databank Searching,” C. A. Orengo, N. P. Brown & W. R. Taylor, *Proteins* **14** 139-167 (1992).

motifs
database

Searching the Motif Database

The Motif Database utility is based on the same principles as the Superpose Motif utility and the theory is explained in that chapter. The motif data for proteins from the Brookhaven Protein Databank are stored in the file \$HYD_LIB/motif.geo. The reference template which is used to search the database is taken from a currently selected MSF. The template can be the entire MSF structure or a fragment specified by deselecting secondary

structure elements using the **Change Active** tool. The time taken for a motif search job to run is dependent on the size of the search fragment and it is worthwhile to ensure you have defined the minimal reference motif. The search job can be left to run and the results from the log file reviewed later.

Tools and Options

Using this utility, a search can be done to match secondary structures in a test structure to known references. The reference database used is \$HYD_LIB/motif.geo. The **Auto Database Search** can also be used in non-graphical QUANTA in a batch file.

There can be only one molecule active when using this module. If there is more than one active molecule when the module is entered then all but the first are made inactive. Additional structures can be appended to the database.

The motif database can be used in two ways:

- Finding a match with a specific structure
- Running an automatic search against all the entries

An automatic database search generates a log file that can be browsed for the results of the database scan.

Change Active

Initially all secondary structure elements are active and the vectors are displayed with fat lines. Selecting elements switches them off and on. When this option is selected, the **Pick Element** tool is selected by default and the other tools are ungrayed. When this option is selected again all the tools are switched off and grayed.

Pick Element

This option is grayed until the **Change Activity** option is selected. It allows you to deselect any secondary structure element of the active structure. The tool is switched off when either the **Pick Element Range**, **Pick Domain**, or **Change Active** is selected.

Pick Element Range

This option is grayed until the **Change Activity** option is selected. It allows you to pick a range between two secondary structure elements and deselect all the elements within the range. If there is some conflict, for example if not all elements in the range are the same state, you are prompted with a dialog box to further specify conditions. The tool is switched off when either the **Pick Element**, **Pick Domain** or **Change Active** is selected.

Pick Domain

This option is grayed until the **Change Activity** option is selected. It allows you to deselect any defined domain. The tool is switched off when either the **Pick Element**, **Pick Element Range** or **Change Active** is selected.

Overlay Database

If no database file has been selected, this option opens the File Librarian, from which a database file may be chosen. Once a database has been selected, you are presented with a scrolling list of all structures in the database from which to choose. The selected structure is then tested against the reference motif defined by the secondary structure elements in the active molecule. Vectors are drawn representing the secondary structures of all the matches superposed on the reference motif. The browsing tools are ungrayed.

Auto Database Search

If no database file is open, this tool prompts you to select one and enter the name of a log file. Every structure in the database is searched for the reference motif defined by the currently active secondary structure elements. Depending on the size of the database, this computation may take considerable time. As each structure is searched in the database, its name is listed below the command line. The final results are listed to the textport and a log file in the same format as the tables.

Table Database Results

This option reads the logfile created by the **Auto Database Search** tool. It displays the information about the motif searches for the reference molecule, and displays the possible available overlays after a search is done. There are two tables displayed, **Overlay Motif** and **Secondary Structure Elements** tables.

The Overlay Motif table displays information for each matched overlay. The columns from left to right are: the overlay number, the overlay name from the database, the number of elements matched, and the RMS difference of the superposition. The subsequent columns identify elements that are matched

The Secondary Structure Elements table displays information on each secondary type in the active molecule. The columns from left to right are: molecule name, element number, secondary structure type, the ID of the first residue in the secondary structure, and the ID of the last residue in the secondary structure.

Show Overlay

This tool displays the matched overlays and unmarks the browse tools. After a search is completed and the **Table Database Results** tool has read the search results, you can display the results.

The Display Overlay From Database Search dialog box opens. One database structure must be selected. Since the atomic coordinates of the database structure are not known at this stage, only the secondary structure vectors can be displayed.

Options...

This tool opens the Motif Superposition dialog box, with options for different parameters. These criteria are used in matching secondary structure elements, in specifying cut-off values and in choosing the number of matches to display. For the matching secondary structure options, toggles are set for each option determining which is used and the respective cut-off values.

18.. Motif Database

All Overlays

This tool is masked until a search is performed and the **Show Overlays** tool has been picked. This is the default browse option that displays all resultant overlays.

Next Overlay

This tool is masked until a search is performed and the **Show Overlays** tool has been picked. This tool displays each motif in ascending order. Once the last motif is displayed, the tool sequences back to the first.

Previous Overlay

This tool is masked until a search is performed and the **Show Overlays** tool has been selected. This tool sequentially displays each motif in descending order. Once the first motif is displayed it sequences back to the last.

Select Overlay(s)...

This tool is masked until a search is performed and the **Show Overlays** tool has been selected. From a toggle list you select one or more overlays can be selected for display.

Clear Display

This tool is masked until a search is performed and the **Show Overlays** tool has been selected. It removes from the display any overlays and masks the browse tools.

Select Database...

This tool opens a File Librarian from which you can select or change the database file used.

Save Structure to Database

This tool saves the active molecule to database. If no database file is open you are prompted to select one. This tool enables users to generate their own database or extend existing databases.

Finish

If molecules have been superposed but not saved this tool prompts the to save coordinates to an MSF file.

Motif Database Log File

This section describes the contents of a Motif Database log file, which can be browsed using the **Sys Window** option from the QUANTA **File** menu. The file uses the following keywords:

**MOLECULE msf_name*

The name of the molecule MSF from which the reference motif was generated in msf_name.

**LIBRARY library_file*

The name of the library file is library_file.

**TEST molecule_name*

The test structure is molecule_name, which, by default, is a four-letter PDB code.

**NMATCH number_of_matches*

The number of matches for molecule_name is number_of_matches. This line is followed by number_of_matches lines, one for each match, and uses the format:

```
N number_of_elements score list_of_matches
```

where:

N

Is the match number for that test molecule;

NUMBER_OF_ELEMENTS

Is the number of elements which are matched;

SCORE

Is the score from RMS superposition of matched element endpoints;

LIST_OF_MATCHES

Is a list containing the number of fields equal to the number of elements in the reference structure. The nth field contains the number of the element in the test structure that matches the nth element in the reference structure. If there is no match to the nth element of the reference structure, the field is blank.

19. Protein Health

Overview

The Protein Health utility can be accessed as a separate application under the **Application** pulldown or it can be accessed from within the Protein Design application via the tool in the **Protein Utilities** palette. Protein Health will identify features in a protein structure which are wrong, for example the wrong chirality of C α atoms, or uncommon, and therefore worthy of closer examination, for example main chain conformations that fall outside accepted regions on the Ramacandran map, or side chain conformations that do not correspond to regularly observed rotamers. This application can be used as an aid to model building and provide criteria to judge the quality of imported data such as PDB files.

Multiple conformations of the same structure can be read from a csr file and the results for all conformations tabulated for easy comparison.

This chapter describes:

- Using Protein Health
- Analysis of Multiple Conformations
- Tools and Options
- Using the Health tools

For more information see:

Protein User's Reference:

- Accessibility
- Model Side Chains

References

C. M. Wilmott & J. M. Thornton, *Prot. Eng.* **3** 479-493 (1990).

Ponders & Richards, *J.Mol. Biol.* **194** 775-791 (1987).

Sutcliffe et.al., *Prot. Eng.* **1** 385-392 (1987).

R. L. Dunbrack & M. Karplus, *J. Mol. Biol.* (in press).

Using Protein Health

This palette is activated from the Protein Utilities or the QUANTA **Applications** menu. These tools are intended to provide a guide when model

building or to enable judgements to be made on the quality of imported data, such as PDB files. The tools in this palette will identify:

- Atoms with undefined coordinates
- Main chain conformation outside accepted regions on the Ramachandran map
- Non-planar peptide bonds
- Side chain conformation not corresponding to commonly observed rotamers
- Buried polar atoms which are not hydrogen bonded
- Hydrophobic and hydrophilic residues in inappropriate environments
- Atoms which are too close
- Holes inside the structure

There are three means of presenting the results of the health check: listing to a file, listing to the textport, or by highlighting the bad features on the molecule display. If multiple conformations are read in from csr files then a comparison of their health properties is presented in a table.

If the **Display Exception** tool is active, then the molecule will be colored to indicate health exceptions. The atoms involved in the exception will be colored — for example, the side chain atoms are colored when the side chain is not close to a rotamer optimal conformation. It may simplify the display if you use the **Display C α and Side Chain** option from the **Molecule Colors** tool on the **Protein Utilities** palette. The one health exception which may be overlooked with this simplified display is a buried, not hydrogen bonded, backbone N or O atom.

Analysis of Multiple Conformations

Several model building methods will generate multiple alternative models; particularly NMR experiment or automated model building using **MODELER** will generate multiple models. Analysis of the protein health criteria for each model can indicate which regions of the model may be poor and comparison of health exceptions between the models can suggest which are the better models.

In the **QUANTA** environment, multiple models are usually stored in a csr file and the **Tabulate MultiConformations** tool will read the confrontations from a csr file and apply selected health checks to all conformations. Csr files can be generated from PDB files (such as are output by **XPLOR** NMR refinement) using the tool in Protein Health or they can be generated in the Protein **MODELER** application. Csr files contain multiple sets of coordinates but they must be associated with an MSF which contains other

data such as atom names and residue types. The **Tabulate MultiConformations** tool requires that the MSF associated with the csr file is the only selected molecule. If this is *not* the case, then the required MSF will be selected automatically. The health checks which are currently active and highlighted on the palette are applied to each conformation in the csr file. As the checks are performed, the molecule window is updated to display the current conformation and, if the **Display Exception** tool is active, colors them to indicate health check exceptions. The **Legend** tool on the **Protein Utilities** palette should be activated to provide a key to the color coding.

When the analysis is complete the full results are presented in a table. The molecule conformation reverts to that from the MSF file but the **Display Conformation** tool becomes available for you to display any of the conformations.

The multi-conformation table has one column per conformation and one row per health exception. The rows are labelled with the name of the residue and a short mnemonic for the exception:

- **MAIN** the mainchain phi/psi angles are not in a good region of the Ramachandran plot
- **SIDE** the sidechain is not within a toleration limit of an optimal rotamer
- **BURPOL** there is a buried polar atom which is not hydrogen bonded.
- **ENVIR** the residue side chain is in an inappropriate environment - either an hydrophobic residue exposed to solvent or an hydrophilic residue buried in the protein interior.

If a particular conformation has a health exception then the appropriate cell in the table is marked with an “X”. At the top of the table the number of exceptions for each conformation is given. The left-most column in the table has the number of conformations which have the exception.

If either the **Close Contacts** or **Holes** option are active then the number of close contacts or holes in each conformation are listed at the top of the table. You will need to use the **Display Conformation** tool to display the close contacts or holes for each individual conformation.

To see the exceptions reported in the table you can pick the table to reset the display. You can change the displayed conformation by picking the name of the required conformation from the top row of the conformation table. If you pick a residue ID from the left hand column of the table, then the display will be centered on that residue in the currently displayed conformation. Picking a cell in the body of the table will display the conformation given by the table column and center the display on the residue given by the table row.

Change display by picking conformation table

Tools and Options

Undefined coordinates

This tool identifies two classes of undefined coordinates:

- Main chains: colored in red
- Side chains where the main chain is defined but one or more atoms in the side chain are undefined: colored in blue

Boxes are drawn and labelled with the residue ID for each residue that has undefined coordinates. This information is also listed to the textport.

Mainchain conformation

The peptide bond is normally expected to be planar, such that the omega angle, $C\alpha-C-N-C\alpha$, is about 180° . By default omega angles less than -170° or greater than 170° are flagged.

The data file \$HYD_LIB/protein_param.dat contains a Ramachandran map. This map is derived from analysis of observed conformations in well-resolved crystal structures*. The Ramachandran map is divided into a grid of 10° by 10° blocks. The commonly observed conformations lie in six different regions on this map. The health check flags any residue with a conformation which lies outside these regions.

Sidechain conformation

Several analyses of the protein databank show that most sidechains occur in a limited range of conformations. These observed conformations are usually called rotamers.

Several libraries of the commonly occurring rotamers are described in the literature and three of them are used within QUANTA. The libraries differ in the form of the analysis which was applied, particularly in the handling of the dependence of the side chain conformation on the backbone conformation.

These rotamer libraries are available to use with either protein health checking or in side chain model building. They are listed in the file \$HYD_LIB/protein_param.dat. except for the Dunbrack and Karplus rotamers which are listed in \$HYD_LIB/harvard_torsion.dat.

The Dunbrack and Karplus Rotamer Library

This library was generated by compiling statistics on the side chain torsions observed for a given range of main chain torsions. The side chain torsion space was divided into chemically sensible rotamers. For example, the torsion about a bond connecting two sp^3 carbon atoms is split into 3 rotamers:

gauche $+ 0^\circ < \chi < 120^\circ$

trans $120^\circ < \chi < 240^\circ$

gauche $-(minus) -120^\circ < \chi < 0^\circ$

*C.M. Wilmott & J.M. Thornton (1990). Prot. Eng. 3 479-493.

Table 2. Rotamer Library Types

Rotamer Libraries	Description
Ponder and Richards	No analysis on main chain conformation dependency. The percentage of observed occurrence is given for each rotamer.
Sutcliffe et.al.	Rotamers are designated as specific for helix or strand or independent of main chain conformation. No percentage of observed occurrences is listed.
Dunbrack and Karplus	In analyzing dependency on main chain conformation, the torsion space for phi and psi was divided into a grid of 20° by 20° blocks.

If the observed torsion falls within the appropriate range, it is counted in the statistics even if it is a long way from the presumed optimal center of the range.

Dunbrack and Karplus rotamer definitions

Table 3. Recognized rotamers for chi1 (all residue types)

rotamer	chi1	definition
1	g+	0° < chi1 < 120°
2	t	120° < chi1 < 240°
3	g-	-120° < chi1 < 0°

For the Karplus rotamer, any sidechain not within any accepted rotamer range for its main chain conformation is assigned color 7 (green). If the side chain is within an acceptable rotamer conformation, but has a torsion angle more than the cutoff (default of 30°) from the center of the rotamer range, it is assigned color 1 (light green).

Chirality

The chiral atoms within the standard amino acids which are tested are listed in [Table 8](#). For the valine and leucine sidechains, bad chirality results from inappropriate atom naming.

Buried Polar Atoms

This tool indicates buried polar atoms for all oxygen or nitrogen atoms that have a solvent accessibility less than a given cutoff (default of 0.01) and are not hydrogen bonded.

The hydrogen bonds used in this analysis are derived using generous criteria which includes the “near” hydrogen bonds as defined by the parameters in the **Hydrogen Bond** utility.

Table 4. Recognized rotamers for chi2
Residue types: Leu, Ile, Gln, Glu, Met, Arg, Lys, and Pro^a

rotamer	chi1	chi2
1	g ⁺	g ⁺
2	g ⁺	t
3	g ⁺	g ⁻
4	t	g ⁺
5	t	t
6	t	g ⁻
7	g ⁻	g ⁺
8	g ⁻	t
9	g ⁻	g ⁻

^aThe values for g⁺, t, and g⁻ remain constant.

Table 5. His, Phe, and Tyr (NB by definition all chi2 > 0)

rotamer	chi1	chi2		
1	g ⁺	0°	< chi2	< 60°
2	g ⁺	60°	< chi2	< 120°
3	g ⁺	120°	< chi2	< 180°
4	t	0°	< chi2	< 60°
5	t	60°	< chi2	< 120°
6	t	120°	< chi2	< 180°
7	g ⁻	0°	< chi2	< 60°
8	g ⁻	60°	< chi2	< 120°
9	g ⁻	120°	< chi2	< 180°

Table 6. Residue type Trp

rotamer	chi1	chi2			
1	g ⁺	0°	< chi2	< 180°	(+90° rotamer)
3	g ⁺	-180°	< chi2	< 0°	(-90° rotamer)
4	t	0°	< chi2	< 180°	(+90° rotamer)
6	t	-180°	< chi2	< 0°	(-90° rotamer)
7	g ⁻	0°	< chi2	< 180°	(+90° rotamer)
9	g ⁻	-180°	< chi2	< 0°	(-90° rotamer)

Table 7. Residue type: Asp, and Asn

rotamer	chi1	chi2			
1	g+	-90°	< chi2	< -30°	(g- rotamer)
2	g+	-30°	< chi2	< 30°	(t rotamer)
3	g+	30°	< chi2	< 90°	(g+ rotamer)
4	t	-90°	< chi2	< -30°	(g- rotamer)
5	t	-30°	< chi2	< 30°	(t rotamer)
6	t	30°	< chi2	< 90°	(g+ rotamer)
7	g-	-90°	< chi2	< -30°	(g- rotamer)
8	g-	-30°	< chi2	< 30°	(t rotamer)
9	g-	30°	< chi2	< 90°	(g+ rotamer)

Table 8. Atoms tested for Chirality

atom name	residue type
CA	All amino acids
CB	Ile
CB	Val
CG	Leu

Buried hydrophilic and Exposed hydrophobic

Hydrophilic residues normally occur on the protein surface and hydrophobic residues occur in the protein core. Residues often occur in inappropriate environments for several reasons:

- Buried hydrophilic groups usually have their polar atoms involved in salt bridges or hydrogen bonds
- The polar atoms can be exposed even if the rest of the side chain is buried.
- Areas of exposed hydrophobic residues occur on the protein surfaces that normally form part of the interface to other proteins or ligands.

The parameter used for residue accessibility is the side chain fractional accessibility. This is the sum of the accessibilities for the side chain atoms, relative to the maximum possible accessibility for the side chain. This assumes an extended conformation and minimal occlusion of the side chain by the neighboring main chain. Hydrophobic residues with greater than 0.9 accessibility and hydrophilic residues with less than 0.1 accessibility are flagged.

Close Contacts

This tool indicates close contacts when the distance between atoms are less than some proportion of the sum of their van der Waals radii (default of 0.80). Close contacts within one residue, usually due to bad side chain con-

19.. Protein Health

formation, and contacts between C β atoms and neighboring residues, are not flagged.

Holes

This tool checks for holes within protein structures that are large enough to accommodate a solvent molecule. The tool identifies probable solvent sites, and is potentially useful to crystallographers trying to identify solvent molecules in electron density refinement. The minimum radius of the hole used can be set by the **Options** tool.

The solvent is modeled as a simple sphere of radius about 1.3 angstrom. This method considers the protein on a 3D grid and marks each grid point as either protein or solvent, depending on its proximity to a protein atom. A point is considered as 'protein' if it is within the atomic radius plus the solvent radius of the center of a protein atom. A flood-fill technique is then used to identify all the connected solvent points that constitute the bulk solvent around the protein. The remaining solvent grid points, not connected to the bulk solvent, are putative holes within the protein. A second pass analysis at a higher resolution identifies if there is truly enough space for the solvent sphere to fit. The centroid and volume of each hole is listed to textport. The volume given is the space within which the centroid of a solvent sphere could move without the solvent sphere contacting the neighboring atoms. The reported volume will be greater than 0.0 but may be a very small value.

Options

This tool opens the Protein Health Options dialog box from which you can change default variables for the Protein Health tools.

Select Active Residues...

This tool activates the **Residue Selection** palettes. Whilst this tool is active the health checks will be applied to only the selected residues.

Display Exceptions

This tool, which is active by default, displays the exceptions to the currently active health checks on the molecule. The molecule is colored gray, color 11, and areas of bad structure are highlighted in bright colors. The legend on the bottom left of the screen describes what each color indicates. When the **Close Contacts** tool is selected, close atoms are indicated by dashed lines.

List Exceptions to textport

When this tool is picked the exceptions to the currently active health checks are listed to the textport.

Write Exceptions to File

When this tool is picked the exceptions to the currently active health checks are written to a file with the name *molecul_health.out*. If close contacts are active then they are written in the file *molecule_bumps.out*.

Phi/Psi Plots

This tool generates a phi/psi plot with the allowed regions drawn in the colors listed in [Table 9](#). The selected residues are marked on the plot.

Phi/psi Options

This opens the Define phi/psi Dihedrals dialog box that enables you to change the default torsions angles.

Table 9. Allowed Main Chain Conformations^a

Conformation	Code Letter	Color on Plot
α_R Right hand α helix	A	8 purple
β_E Extended β strand	B	4 yellow
β_P Poly proline β strand	P	13 brown
α_L Left hand α helix	L	12 pale blue
γ Gly left handed α helix	G	10 salmon pink
ϵ Gly assessable region	A	5 white

^aWilmot & Thornton, 1990, Protein Engineering, 3, 479-493.

Highlights

This tool enables cross-highlighting of picked residues on the phi/psi plot, the sequence viewer or structure.

List Phi/psi to Textport

This tool writes the Phi/psi angles and the side chain angles to the textport.

Write Phi/psi to File

This tool writes the Phi/psi angles and the side chain angles to a file molecule_torsions.out.

Convert PDB files to CSR

You are prompted by the file librarian to select an XPLOR PDB file and to give a name for the csr file which is created.

Tabulate MultiConformations

You are prompted to select a csr file. If the MSF associated with the csr file is not the only currently selected MSF then a reselection is performed automatically. The currently selected, highlighted health tools are applied to every conformation in the csr file and the results tabulated. The Display Conformation tool becomes available. If the **Tabulate MultiConformations** tool is picked again the current table will be overwritten.

Display Conformation

You can select a conformation number to display or opt to return to the conformation of the MSF. The conformation is displayed with the health exceptions that are listed in the table. Other health checks can be performed on this conformation by picking the appropriate tool.

20. Using MODELER

Overview

MODELER is an automated protein homology modeling package that builds a 3D model of a protein sequence using data from the alignment of the sequence with one or more homologous structures.

This chapter describes:

- MODELER
- Accessing MODELER
- Displaying MODELER results
- Evaluating MODELER results
- Deleting files after a MODELER run
- MODELER files and file modifications
- The MODELER control file
- Modifying files to modify models

References

Introduction to comparison of protein structure

Sali, A. Blundell, T. L., “Definition of general topological equivalence in protein structures: A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming,” *J. Mol. Biol.*, **212** 403-428 (1990).

Method

Sali, A. and Blundell, T. L., “Comparative protein modeling by satisfaction of spatial restraints,” *Mol. Biol.*, **234** 779-815 (1993).

Application

Sali, A., Matsumoto, R., McNeil, H. P., Karplus, M., and Stevens, R. L., “Three-dimensional models of four mouse mast cell chymases. Identification of proteoglycan-binding regions and protease-specific antigenic epitopes,” *J. Biol. Chem.*, **268** 9023-9034 (1993).

Sali, A and Overington, J.P., “Derivation of rules for comparative protein modeling from a database of protein structure alignments,” *Protein Sci.*, **31** 582-1596 (1994).

using
modeler

MODELER

MODELER derives spatial restraints such as inter-C α distances and dihedral angles from each homologous structure. The weight assigned to the derived restraints reflect the degree of local sequence similarity between the homolog and the model sequence. The restraints are expressed as probability density functions (pdfs) that describe a model structure. The model is derived by minimizing violations of the restraints. The optimization process to generate the model consists of applying the variable target function as well as conjugate gradients and running molecular dynamics.

The **MODELER** program was developed by Dr. Andrej Šali at Birkbeck College (London), Imperial Cancer Research Fund (London) and Harvard University. It is described in detail in papers listed in the references section.¹⁻⁴

QUANTA provides a simple interface that you can use to set up and run a **MODELER** job. The jobs typically take between five minutes and several hours to run so they are generally run in the background. Model structures generated from a **MODELER** run can be read back into QUANTA for display.

This chapter describes the QUANTA/**MODELER** interface, steps to take to use **MODELER** from within QUANTA, and functionality that you can access. It covers the following areas:

- Accessing **MODELER** (page 139)
- Displaying **MODELER** results (page 141)
- Deleting files after a **MODELER** run (page 143)
- **MODELER** files and file modifications (page 143)

You can access **MODELER** functionality that is not currently interfaced with QUANTA by editing the control file generated within QUANTA and then running a **MODELER** job. The full **MODELER** functionality and interface is described in the **MODELER** User's Guide and Programmer's Manual. This book is available on-line as a postscript file in the following location:

`$MODELLER_ROOT/doc/manual.l.ps.Z`

Accessing MODELER

The tools to **Run MODELER** and **Read MODELER Results** are available in the Protein Design application. These tools and additional tools to analyze the MODELER results are available in the Protein MODELER application.

Before you begin

Before running **MODELER**, you should align and superpose the sequence and homologous structures using the Align and Superpose utility which is available in either the Protein Design or Protein MODELER application within QUANTA. For more information, see Chapter 5. Alternatively alignments generated outside QUANTA can be read in using the Sequence utility on the **Files** pulldown and selecting **Read Sequence/Alignment File**. You will have the option to read alignment files in several commonly used formats.

MODELER does not require multiple homologous structures to be superposed. However, superposition may help you in studying the structures. In contrast, the alignment of the structures is the most significant input to **MODELER**. If there is any uncertainty in the alignment of the sequences you are studying, it is worth your time to try all alignment possibilities. You can improve alignment decisions by checking the final pdf value of each **MODELER** run to determine the residues that have poor restraint satisfaction, then adjusting the alignment accordingly.

You can use fragments to model limited segments of structures by aligning the fragment to the appropriate part of the model sequence. This process can be useful if a motif from a non-homologous structure is a good analogue for part of the model.

You can build models in the presence of ligands or prosthetic groups but this requires manual generation of restraints to position the ligand.

To start MODELER

1. Select **Applications** from the main menu, then select **Protein MODELER** from the **Applications** pulldown. The **Protein MODELER** and **Protein Utilities** palettes are displayed.
2. Select **Run Modeler** in the **Protein Design** palette. A dialog box is displayed.
3. Specify run parameters as described in the following paragraphs:
 - Select the sequence for a target model.

Currently active sequences are listed near the top of the dialog box. Choose one of these as the sequence for which a model will be built.

This sequence can come from either an MSF file or a sequence file. If there are already coordinates for the sequence, they will be completely

ignored by the interface. All other active molecules will be used as homologues for the model.

- In the **Root name for generated models** data entry field, enter a name for the model to be built.

The name should be unique because it will be used as the root name for files and a directory created in running the **MODELER** job. If the name is not unique, you will be prompted to enter an alternative.

- In the **Number of models** data entry field, enter the number of models you want built.

By default, **MODELER** will build one model, but multiple models can be generated.

- Enter a value for rms deviation among initial models. The default value is 4.0 Å between C α of the models.

An initial model is generated by copying and averaging the homologs. Multiple models are generated by perturbing this initial template to generate models with a given RMS deviation in coordinates. All models are then refined independently. Their structures may converge or diverge.

- Select the optimization level during refinement.

Most of the time required for a **MODELER** run is used in optimizing the structure. The optimization protocol involves multiple cycles of refining the model coordinates as the scaling of the restraint parameters is gradually varied to bring in tighter restraints. Available optimization protocols are:

Full — Optimization using copying and averaging of template coordinates, variable target function with conjugate gradients, and molecular dynamics with simulated annealing. Full is the default and recommended selection.

Partial — Optimization using copying and averaging the template coordinates plus variable target function with conjugate gradients.

None — No optimization. Copying and averaging the template coordinates only.

- Choose a run option.

By default, a **MODELER** job starts automatically in background mode on your local host. You can change this by checking the appropriate box:

Prepare control file only; do not run **MODELER**.

The **MODELER** control file, input alignment file, and PDB coordinate files are always written. If you want to modify the control file before

running the job, choose this option, edit the control file, then run the job outside QUANTA by entering the command:

```
modeler model_name
```

Run **MODELER** on remote host.

When you check this box, a Select Host dialog box is displayed. For more information, see “Using External Applications” in *QUANTA Generating and Displaying Molecules*.

MODELER jobs take between five minutes and several hours. Job time increases with the size of the protein and the number of homologs.

To check if a **MODELER** job is complete, select **Show Status** in the **Calculate** pulldown.

Displaying MODELER results

After a **MODELER** job is run, you can display the models within QUANTA.

To display models

1. Select **Display MODELER Results...** in the **Protein MODELER** palette. A **File Librarian** listing the **MODELER** control (.top) files is then displayed.
2. Select a **MODELER** control file (*model_name.top*). QUANTA reads the log file associated with this job to check that the job has run successfully and then reads in the output model files for this job from *model_name_dir/model_name.B9999nn.pdb*, and creates MSF files with the file names *model_name_nn.msf*. The *nn* designation is an incremental number for each model. This number is 01 if there is only one model.
3. Click open. The model structures are displayed in the molecule window. The sequence of the model structures is automatically aligned to the input homologs for the model if they are still selected within QUANTA. Models in MSF files can be manipulated in QUANTA like any other MSF files. The final pdf for the model is listed in the textport.

Evaluating MODELER Results

MODELER generates a large number of restraints to force a model towards the structure of the homolog(s) and to maintain good structure geometry. At the end of a MODELER refinement not all restraints will have been satisfied and the violations associated with each residue can be summed to give

an indication of the quality of the model in and around that residue. There is not a simple interpretation of the absolute value of the restraint violation parameter since parts of a model, for example residues adjacent to an insertion or deletion, are almost certain to have higher restraint violations. The violations are useful in making comparisons between different models. Models with lower violations in a region are probably better in that region. If these models have been generated from different initial alignments of the model sequence and the homologs, then lower restraint violations are a strong indication of better alignment.

In the PDB files output by **MODELER** the **residue restraint violation** is **listed in the place of** the temperature factor parameter. In an MSF file, this information is stored as the fourth parameter (that is, BVALUE).

When the **Display MODELER Results** tool is used the restraint violations are shown as a graph in the sequence viewer. To redisplay the graph use the **Plot Violations** tool.

You can use the **Color Violations** tool to color the molecule and sequence by the violation parameters. The violation values are split into four ranges between the maximum and the minimum values observed for all the residues in the active molecules. The residues are colored white, pale yellow, yellow and red for increasing violations. The **Legend** tool on the **Protein Utilities** palette can be used to display a legend with the key to the color coding.

You can list the restraint violations for each residue by selecting the C α atoms, then typing:

```
LIST ATOM
```

in the command line. The restraint violations are listed in the textport in the column labeled 4th.

Using Protein Health

The tools in Protein Health can be used to indicate improbable features in a model. Protein health exceptions, such as a residue main chain not in a regularly observed conformation or buried polar atom which are not hydrogen bonded, may be indicative of a poor model which, in turn, may indicate a poor choice of homologs or poor alignment. The Protein Health tools can be applied to the MSF files generated by the **Read MODELER Results** tool. If you have several MODELER models it may be easier to write them all to one CSR file using the **Write CSR File** tool and then using the **Tabulate MultiConformations** tool within Protein Health to produce a table which summarizes health exceptions for all of the models (see chapter 19.).

Using Profile Analysis

Profile analysis can be applied to the models generated by MODELER and will plot graphs of scores per residue in the sequence viewer which aid comparison of the quality of models (see Chapter 1.).

Deleting files after a MODELER run

When you begin a **MODELER** job, a directory is created below your current working directory and all files related to the job are placed there except the control and log files.

The name of this directory the same as the name of the control file except for extensions. Models generated by **MODELER** are placed in this directory. Generally, you will read these files through the QUANTA interface.

When you have finished a **MODELER** run, after the final model .pdb files have been read and MSF files have been generated in QUANTA, you can delete the job directory.

To delete a MODELER job directory

Type the following on the command line:

```
sys rm -rt directory_name
```

This will remove the directory and the **MODELER** files in it.

MODELER files and file modifications

This section provides information on the files used by QUANTA in running **MODELER**. None of this information is necessary for you to use **MODELER** from the QUANTA interface.

When a **MODELER** job is started, the following files and directory are created by QUANTA:

- **model_name.top** — A control file in the current working directory.
- **model_name_dir** — A directory below the current working directory that contains input, output, and intermediate files.
- **model_name_dir/model_name.aln** — An alignment file in the usual QUANTA alignment file format.
- **model_name_dir/molecule_name.pdb** — A PDB- format file of a homologous structure coordinates.

MODELER creates the following files:

- **model_name.log** — A log file of the **MODELER** job in the working directory.
- **model_name_dir/model_name.ini** — An initial model from copying and averaging homolog coordinates in PDB format.

- `model_name_dir/molecule_name_.dih.Z`, `model_name_dir/molecule_name_.psa.Z`, and `model_name_dir/molecule_name_.ngh.Z` — Three scratch files containing dihedral, accessibility, and residue neighbor analysis for each input homologous structure.
- `model_name_dir/model_name.rsr` — A constraint file derived for the model structure. The format is described in Part II of the **MODELER User's Guide and Programmer's Manual**.
- `model_name_dir/model_name.sch` — A schedule file that defines the cycles of optimization to be applied. The format of the file is:

Column 1 — An index to the schedule step.

Column 2 — Optimization method (1=CG, 2=Newton Raphson 3=MD).

Column 3 — For the variable target function method restraints, the maximum number of residues and the span of the restraints.

Column 4 and beyond — Scaling factors for the standard deviations of restraints, one column per physical restraint type. The types are listed in `$MODELER_ROOT/modlib/feats.lib` (currently, 1-24 types).
- `model_name_dir/model_name.D0000nn` — A file containing the optimization trace, showing the value of the objective function, gradient, temperature (MD only), and kinetic energy for each cycle in the optimization.
- `model_name_dir/model_name.B9999nn.pdb` — PDB format coordinate files for final models. If there is only one model built, *nn* will be 01. If more than one model is built, *nn* will be 01, 02, 03...

The MODELER control file

MODELER works as a command interpreter for a command language called TOPS. The TOPS language is similar to FORTRAN and includes commands to define variables, open and close files, call subroutines, and do some flow control.

The language is described in Chapter 3 of the **MODELER User's Guide and Programmer's Manual**. A **MODELER** command file is written in the TOPS language and only when a valid **MODELER** command is recognized by TOPS is the appropriate **MODELER** function run. The TOPS language provides the framework for flow control for running **MODELER**.

MODELER commands are described fully in Chapter 2 of the **MODELER** User's Guide and Programmer's Manual along with the keywords that allow you to change the parameters for the commands. A full list of commands and keywords is in the file \$MODELER_ROOT/modlib/top1.ini. This file is read every time **MODELER** is run.

A complete **MODELER** run requires a lengthy command script. Scripts to perform common functions are provided in the \$MODELER_ROOT/exec directory.

A simple **MODELER** control file for homology modeling consists of commands to specify the input files and set any non-default parameters, and to then call the homol subroutine that controls the model building procedure. For example, the control file written by QUANTA to build a model named dfr_model from two structures, 3dfr and 4dfr is set up in the following form:

Set up the initial defaults:

```
INCLUDE
```

Set the **MODELER** working directory for input, coordinate and output files:

```
SET DIRECTORY = 'dfr_model_dir'
SET ATOM_FILES_DIRECTORY = 'dfr_model_dir'
SET OUTPUT_DIRECTORY = 'dfr_model_dir'
```

Set an extension PDB for the output coordinate files:

```
SET PDB_EXT = '.pdb'
```

Set the format for the input alignment file to QUANTA format and specify the alignment file:

```
SET ALIGNMENT_FORMAT = 'QUANTA'
SET ALNFILE = 'dfr_model.aln'
```

Declare the sequence for which to build models. The sequence is taken from the alignment file and no further information is needed:

```
SET SEQUENCE = 'dfr_model'
```

Declare what molecules in the alignment file are homologues of a known structure. The molecule names must correspond to those in the alignment file. There should be .pdb coordinate files with the corresponding molecule_name.pdb:

```
SET KNOWNS = '3dfr' '4dfr'
```

Next is a comment line which is not interpreted by **MODELER**. It is read by QUANTA and used to put the model in the correct sequence alignment to the homologous structures:

```
# QUANTA input sequence = dfr.msf
```

Set the optimization protocol:

```
SET MD_LEVEL = 'refine3'
```

Call the main **MODELER** script from \$MODELER_ROOT/exec to control the program flow:

```
CALL ROUTINE = 'model'
```

Modifying files to modify models

This section describes several file modifications, including adding a disulfide bond and representing hydrogens, that you can make if you want to modify the models that **MODELER** generates. These modifications are not required to complete a **MODELER** run.

To add a disulfide bond

If the target sequence has a pair of cysteine residues aligned to a pair of cysteine residues that form a disulfide bridge in any of the homologous structures, then a disulfide bond will be expected in the model structure and the necessary constraints will be defined automatically. Anticipated disulfide bonds are reported in the log file.

However, if you expect a disulfide bond in the model structure without an analogous bond in the homologous structures, then you must specify this bond in the control file for the necessary constraints to be set up.

MODELER control scripts include a dummy routine called `special_patches` that contains nothing. You redefine this routine to contain the information needed by **MODELER** to generate a disulfide bond.

To redefine the routine, edit the control file to include the following lines after the first `INCLUDE` statement:

```
SUBROUTINE ROUTINE = 'special_patches'  
PATCH RESIDUE_TYPE = 'DISU', RESIDUE_IDS = n1 n2  
RETURN  
END_SUBROUTINE
```

where `n1` and `n2` are the index number of the disulfide bonded residues. Note that the index number is not necessarily the same as the residue ID unless the residues are numbered consecutively from 1 for the first residue.

To represent hydrogens

By default, **MODELER** generates models without hydrogen atoms. By using alternative topology library models, you can generate models with polar hydrogens or all hydrogens.

The default topology is set in the file \$MODELLER_ROOT/exec/___defs.top. This file can be edited to change the default.

Alternatively the following lines can be edited into the control file to setup the addition of polar hydrogens

```
SET HYDROGEN_IO = 'ON'
SET TOPLIB = '${LIB}/top_polh.lib'
SET TOPOLOGY_MODEL = 2
SET PARLIB = '${LIB}/par.lib'
```

Or, to use an all-hydrogen representation, edit the control file:

```
SET HYDROGEN_IO = 'ON'
SET TOPLIB = '${LIB}/top_allh.lib'
SET TOPOLOGY_MODEL = 1
SET PARLIB = '${LIB}/par.lib'
```

Multiple segment models

MODELER will generate models for multi-segment (i.e., multi-chain) proteins without any user intervention if the input alignment file has the segment delimiters, “/”. The alignment files generated by QUANTA will have the segment delimiters for sequences taken from MSF files but since the QUANTA sequence-only data has no segment information there will be no segment delimiters in the alignment file. If your model structure is a multiple segment protein and its sequence is in the form of sequence-only data then you should do one of two things:

- Generate the MODELER input files and then manually edit the alignment file to insert “/” segment delimiters and then run MODELER outside QUANTA.
- Generate an MSF for the sequence using the tool in the Protein Editor and then use the Atom Property Editor to enter the segment information (as described in Chapter 4).

Defining MODELER Restraints

The Constraints Editor

The CHARMM Constraints Editor within QUANTA can be used to define MODELER restraints and write out an appropriately formatted file or to convert NOE information to MODELER restraint file format.

The appropriate tools can be accessed using **Constraints Options and Dihedral/Distance** in the **CHARMM** pulldown. The module includes options to import various constraint formats and to enter constraints by selection and editing the constraint table.

The MODELER restraint file format is described in the MODELER User Guide. Note that the MODELER restraint file references atoms by their number rather than by an atom and residue name and it is therefore very important to generate the restraint file for the correct molecular structure. The simplest way to ensure that the structure is correct is to run MODELER with the **Optimization level** set to **None** so that an initial model is gener-

ated but a full refinement is not performed. The initial model will be in the MODELER work directory with the name *model_name_dir/model_name.ini* and is in PDB format. This file should be read into QUANTA and used in the Constraint Editor.

In the MODELER restraint file the magnitude of a restraint is specified by the standard deviation in angstroms or radians (SD). This is different from the Charmm convention of specifying a force constant *k*. The relation between these is:

$$k = 0.5 \cdot R \cdot T / SD^2$$

When exporting MODELER restraint files from the Constraint Editor, the necessary conversion is made from the force constant *k* to the standard deviation assuming a temperature of 297.15 K.

Dihedral constraints are defined in the MODELER restraint file as single Gaussian harmonic potentials and the target parameter in the Constraint Editor is taken as the ideal value. NOE distance constraints are interpreted in the MODELER restraint format as two restraints: one for upper and one for lower distance bound. The upper and lower bound distances reported in the Constraints Editor table, and not the target value, are used in defining the MODELER restraints.

Using Additional Restraints in MODELER

Once a file containing user defined restraints is created and named, for example, *user_defined.rsr* can be read by MODELER by inserting the following lines in the control file generated by Quanta after the first line:

```
SUBROUTINE ROUTINE = 'special_restraints'  
READ_RESTRAINTS FILE = 'user_defined.rsr' , ADD_RESTRAINTS = ON  
RETURN  
END_SUBROUTINE
```

special_restraints is a subroutine which is called at those places in the modeler protocol where restraints are set up. By default, the subroutine contains no functionality but you can redefine the subroutine to incorporate such functionality as reading an additional restraints file.

Including Ligands

If MODELER is used to build a protein model containing a ligand, then the ligand is treated as a rigid body. Its conformation is fixed but the position relative to the protein may alter. To anchor the ligand in the correct position distance restraints between ligand and protein atoms can be defined using the protocol outlined above.

MODELER Theory

Protein and nucleic acid sequence determination is now routine in molecular biology laboratories. As a result, the rate of publication of sequence information has increased dramatically. Several organizations have com-

piled databases to centralize the published information and to facilitate its use in research (EMBL database, PIR/NBRF database, GenBank database). On the other hand, structural information from X-ray crystallographic or NMR results is obtained much more slowly. For instance, if the protein is not a simple mutation of one whose structure is already known, it can take months or years to perform a complete structural determination. Because of the disparate rates of sequence and structure determination, there are many proteins for which sequence information is known but the 3D structure is not.

It is advantageous to have a method whereby the conformation of a newly characterized protein can be predicted from its amino acid sequence.* The importance of homology modeling is drastically increased with the explosive multiplication of completely sequenced genomes and parallel structural databases of models based on these genomes.

Early work dealing with building a protein by homology used only one known structure (Browne et al. 1969, Shotton and Watson 1970). Amino acid similarities between the known and unknown proteins were used to determine where one protein would resemble the other. Sequence alignment was done, and the coordinates of the reference protein were used to predict those of the unknown protein.

More structural approaches have been suggested by Greer and Blundell (Greer 1980, 1981, 1985; Blundell 1987, 1988). In their methods, more than one reference protein is used, and a greater emphasis is placed on the conformational similarities between them. Less emphasis is placed on sequence alignment alone as a basis for the model. By determining which portions of the molecules do not vary from one member of a protein family to another, there is greater confidence that extrapolation to a new member will be accurate.

While several reference structures are used in the traditional homology model building process, only one set of coordinates can be used in any one peptide segment. MODELER is able to simultaneously incorporate structural data from one or more reference proteins. Structural features in the reference proteins are used to derive spatial restraints which in turn are used to generate model protein structures using conjugate gradient and simulated annealing optimization procedures.

[Sequence-structure alignment \(Align2D\) \(page 150\)](#)

[Structure alignment \(Align3D\) \(page 152\)](#)

[Building the structure \(page 153\)](#)

[Atom-atom repulsions \(page 157\)](#)

*Comparative or homology modeling is currently the only such method. It is based on the assumption that the structure of the unknown protein is similar to known structures of some reference proteins.

[Distance between two Ca atoms \(page 157\)](#)
[Distance between main-chain N and O atoms \(page 160\)](#)
[Distances between sidechain-sidechain and sidechain-mainchain atoms \(page 160\)](#)
[Main-chain conformation \(page 160\)](#)
[Side-chain conformation restraints \(page 162\)](#)
[Restraining Chi1 side-chain dihedral angles \(page 163\)](#)
[Restraining Chi2 dihedral angles \(page 163\)](#)
[Restraining Chi3 dihedral angles \(page 163\)](#)
[Restraining Chi4 dihedral angles \(page 164\)](#)
[Basis and feature probability density functions \(page 165\)](#)
[Feature PDFs used for restraining model protein features \(page 166\)](#)
[Structure generation using MODELER \(page 169\)](#)
[Modeling loops \(page 170\)](#)
[References \(page 172\)](#)

Sequence-structure alignment (Align2D)

One of the main problems in comparative modeling is the accuracy of the sequence alignment. MODELER extends the global alignment algorithm to align two blocks of sequences (the structure block and the sequence block). The command uses a variable gap-opening penalty which depends on the 3D structure of the sequences in the structure block and tends to place gaps in a better structural context. The variable gap penalty favors gaps in exposed regions, avoids gaps within secondary structure elements, favors gaps after curved parts of the main-chain, and minimizes the distance between the two positions spanning a gap.

The sequences within the two blocks, structure block (first block) and sequence block (second block) should be prealigned and the first block is required to have at least one sequence with known 3D structure. The two blocks of sequences are defined in the MODELER control file as sequences 1 to ALIGN_BLOCK and ALIGN_BLOCK+1 to the last sequence.

The linear gap penalty function for inserting a gap in block 1 of a structure is:

$$g = f_1(H, S, B, C)u + lv \quad \text{Eq. 4}$$

where u and v are the usual gap opening and extension penalties (defined by GAP_PENALTIES_1D) and l is the gap length.

f_1 in the above equation is a function that is at least 1 but can be larger to make gap opening more difficult, as in the following circumstances.

Between two consecutive (i.e., $i, i+1$):

- helical positions
- β -strand positions
- buried positions
- positions where the main-chain is locally straight

The function represented by f_1 is:

$$f_1 = 1 + [\omega_H H_i H_{i+1} + \omega_S S_i S_{i+1} + \omega_B B_i B_{i+1} + \omega_C C_i C_{i+1}]$$

Eq. 5

where:

The weights, ω , are the first four numbers in variable GAP_PENALTIES_2D.

H_i is the fraction of helical residues at position i in block 1.

S_i is the fraction of β -strand residues at position i in block 1.

B_i is the average relative side-chain “buriedness” of residues at position i in block 1.

C_i is the average straightness of residues at position i in block 1.

The original straightness is modified here by assigning a maximal straightness of 1 to all residues in a helix or β -strand.

The linear gap penalty function for opening a gap in block 2 (sequence block) is:

$$g = f_2(H, S, B, C, D)u + lv$$

Eq. 6

f_2 in the above equation is a function that is at least 1 but can be larger to make the gap opening in block 2 more difficult, as in the following circumstances:

1. When the first gap position is aligned with:
 - a helical residue
 - a β -strand residue
 - a buried residue
 - extended mainchain

2. When the whole gap in block 2 is spanned by two residues in block 1 that are far apart in space.

The function represented by f_2 in the above equation is:

$$f_2 = 1 + [\omega_H H_i + \omega_S S_i + \omega_B B_i + \omega_C C_i + \omega_D \sqrt{d - d_o}]$$

Eq. 7

where:

H_i is the fraction of helical residues at position i in block 1.

S_i is the fraction of β -strand residues at position i in block 1.

B_i is the average relative side-chain “buriedness” of residues at position i in block 1.

C_i is the average straightness of residues at position i in block 1.

d is the distance between the two C α atoms spanning the gap, averaged over all structures in block 1.

d_o is a distance small enough to correspond to no increase in the opening gap penalty (e.g., 6 Å).

Parameters ω_H , ω_S , ω_B , ω_C , ω_D , and d_o are specified by GAP_PENALTIES_2D.

For an example, see the standalone Modeler manual at:

<http://salilab.org/modeller/manual/>

Structure alignment (Align3D)

Multiple templates are often used together to model a sequence. The **Align_Structure** command aligns two or more template structures based on their structural similarity. The command first generates a multiple sequence alignment using the progressive pairwise sequence alignment method. Based on the initial sequence alignment, MODELER performs an iterative least-squares superposition of 3D structures. This results in a new multiple structure alignment.

Starting from the initial sequence alignment, the structure alignment involves several cycles, each consisting of an update of the framework and calculation of the new alignment. The new alignment is based on superposition of the structures onto the latest framework. The framework in each cycle is obtained as follows.

1. The initial framework consists of $C\alpha$ atoms in structure 1, but only at the alignment positions where all the structures have a residue. If there is a missing $C\alpha$ atom in any of the residues at a given position, the coordinates for this framework position are approximated by the neighboring coordinates.
2. All other structures are fit to this framework. The final framework for the current cycle is then obtained as an average of all the structures, in their fitted orientations, but only for residue positions that are common to all of them, given the current alignment.

Another result is that all the structures are now superimposed on this framework. Note that the alignment has not been changed yet.

3. The multiple alignment itself is re-derived in $N-1$ dynamic programming runs, where N is the number of structures. This is done as follows.
 - a. Structure 2 is aligned with structure 1, using the intermolecular atom-atom distance matrix, for all $C\alpha$ atoms, as the weight matrix for the dynamic programming run. The dynamic programming is the same as pairwise sequence alignment, where the sequence similarity matrix is replaced by the $C\alpha$ distance matrix.
 - b. Structure 3 is aligned with an average of structures 1 and 2, using the same dynamic programming technique.
 - c. Structure 4 is then aligned with an average of structures 1-3, and so on.

Averages of structures 1- i ($i=2,3,\dots,n$) are calculated for all alignment positions where there is at least one residue in any of the structures 1- i ($i=2,3,\dots,n$). (This is different from a framework which requires that residues from all structures be present.)

Note that in this step, residues out of the current framework may become aligned and the current framework residues may become unaligned. Thus, after the series of $N-1$ dynamic programming runs, a new multiple alignment is obtained. This is then used in the next cycle to obtain the next framework and the next alignment.

The cycles are repeated until there is no change in the number of equivalent positions. This procedure is best viewed as a way to determine the framework regions, not the whole alignment.

Building the structure

Probability density functions (page 154)

“Probability density functions observed in known protein structures (page 154)

Probability density functions

In many areas of molecular modeling, geometric features such as distances and dihedral angles may be restrained by setting lower and upper bounds on their allowed values. A pseudo-energy term and a restraining force are then applied if the geometric feature in a model molecule is found to be outside of these bounds. MODELER describes restraints on such geometric features in terms of a probability density function (or PDF) $p(\chi)$, for a feature χ that is to be restrained. The area under the curve $p(\chi)$ between χ_1 and χ_2 :

$$p(\chi_1 \leq \chi < \chi_2) = \int_{\chi_1}^{\chi_2} p(\chi) d\chi$$

Eq. 8

is the probability of an event in which χ is observed within this range. This must integrate to one over the range of all possible values for the feature χ . This description is more complete than the use of mean values and/or lower and upper bounds as a description for restraints.

In addition, the concept of a conditional probability density function is important in comparative modeling. The PDF used to restrain a geometric feature χ can be written as

$$p(\chi/a, b, c, \dots)$$

Eq. 9

This describes the probability density for χ when the other features $a, b, c, \dots$ are specified. As an example, $p(\chi_1/\text{residue_type}, \phi, \psi)$ could be used to describe the side-chain dihedral angle χ_1 for a known residue type and main-chain dihedral angles ϕ and ψ .

Probability density functions observed in known protein structures

The probability density functions used by MODELER are derived analytically using statistical mechanics and empirically using a database of known protein structures. Stereochemical restraints such as bond lengths, bond angles, dihedral angles, van der Waals repulsions, and disulfide bonds are derived from published parameters for a molecular mechanics forcefield (Brooks et al. 1983). The disulfide bond is restrained by distances, angles,

and the C-S-S-C dihedral angle. Restraints such as those applied to C α -C α distances, N-O distances for main-chain atoms, main-chain dihedral, and side-chain dihedral angle classes are derived empirically from a database of known homologous proteins and their structural alignments.

The method of deriving PDFs from the database of known proteins may be found in Sali and Overington 1994. The database and PDFs depending on it are regularly updated (the last update was July, 1998).

The database is considered a representative sample of globular proteins and is suitable for deriving the relationships between features of protein structures. These relationships are expressed as PDFs. Features are derived at several levels. There are residue-residue relationships such as C α -C α distances, protein-protein relationships such as percentage sequence identity, residue properties such as solvent accessibility, and protein properties such as resolution in X-ray diffraction.

Given the database of proteins, an analysis program is used to derive a table $W_{\chi,a,b,..c}$ that approximates $p(\chi/a, b, c, ...)$:

$$W_{\chi, a, b, ..c} \equiv p(\chi/a, b, c, ...) \quad \text{Eq. 10}$$

This table contains the observed relative frequencies for a feature χ given $a, b, ..c$. Non-linear least squares regression is then used to derive an analytic function that fits the table W .

$$W_{\chi, a, b, ..c} \equiv f(\chi, a, b, ..c, \mathbf{q}) \quad \text{Eq. 11}$$

where $\mathbf{q}$ represents a vector of non-linear fitting parameters. The forms of the probability density function derived for various geometric features used by MODELER are described below.

Probability density functions for bond lengths, bond angles, and dihedrals

A classical harmonic oscillator model for a bond of equilibrium length $\bar{b}$ leads to an energy expression of the form (Sali and Blundell 1993a):

$$E(b) = \frac{1}{2}c(b - \bar{b})^2 \quad \text{Eq. 12}$$

where c is a constant for a given bond type. The probability density function for the bond length is then a Gaussian function:

$$p^b(b) = \frac{1}{\sigma_b \sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{b - \bar{b}}{\sigma_b} \right)^2 \right] = N(\bar{b}, \sigma_b)$$

Eq. 13

where $\sigma_b = \sqrt{kt/c}$ is a standard deviation and only two parameters, $\bar{b}$ and σ_b , are needed to describe this distribution. This functional form is monomodal having a single maximum at $b = \bar{b}$.

The probability density function used to describe a bond angle α is analogous to that described above. The restraint is again a single Gaussian function:

$$p^\alpha(\alpha) = N(\bar{\alpha}, \sigma_\alpha)$$

Eq. 14

The monomodal PDF for a torsion or an improper dihedral γ is

$$p^\gamma(\gamma) = N(\bar{\gamma}, \sigma_\gamma)$$

Eq. 15

This form is used in MODELER to restrain peptide and ring planarities as well as the chiralities of C α atoms and Thr and Ile side-chains. To allow for the *cis*-peptide conformation, the main-chain dihedral ω is restrained to a sum of two Gaussian functions, one centered at 0° and the other at 180°.

$$p^\omega = \omega_1 N(180^\circ, \sigma_\omega) + \omega_2 N(0^\circ, \sigma_\omega)$$

Eq. 16

where ω_1 and ω_2 are fractional weights and their sum equals one. In general terms, a PDF for a bond length, angle, or dihedral f would be described by a functional form (Sali and Blundell 1993a):

$$p = \sum_{i=1}^n \omega_i p_i = \sum_{i=1}^n \omega_i \frac{1}{\sigma_i \sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{f - \bar{f}_i}{\sigma_i} \right)^2 \right]$$

Eq. 17

The mean values and standard deviations used for bond lengths, bond angles, and dihedral angles were obtained from the CHARMM22 parameter set (Brooks et al. 1983, Nilsson and Karplus 1986).

Atom-atom repulsions

The atom-atom repulsion is not described by a harmonic model but uses a PDF of the form:

$$p^v(d) = c \begin{cases} N(d_o, \sigma_w); d \leq d_o \\ \frac{1}{\sigma_w \sqrt{2\pi}}; d_o < d < d_{max} \end{cases}$$

Eq. 18

where d is the distance between the two atoms, d_o is the sum of their van der Waals radii, σ_w is the standard deviation part of the PDF (typically 0.05 Å), and d_{max} is the maximum possible linear dimension for a protein. The constant c is chosen so that this function integrates to 1. This form is flat for distances greater than d_o but presents a soft sphere repulsion for atoms closer than distance d_o .

Distance between two C α atoms

The unknown feature used by MODELER is the difference between two equivalent C α -C α distances, or $d'-d$ where d' is a C α -C α distance in the reference protein and d is that in the model structure. The database of known protein structures is used to find the distribution of $d'-d$ as a function of four variables.

1. The corresponding C α -C α distance (d') in the known structure.
2. The fractional sequence identity (i) of the two aligned sequences.
3. The average of the fractional solvent accessibilities ($\bar{a}'$) of the two residues spanning d' in the known protein structure.
4. The average distance from a gap ($\bar{g}$) of the residues spanning the distance d .

The database and the associated software were used to show that the C α -C α distance difference ($d'-d$) can be approximately modeled as a Gaussian function with a mean of zero and a standard deviation $\sigma(\bar{g}, i, \bar{a}', d')$ that depends upon the values of the four independent variables.

The form of the expression describing $\sigma(\bar{g}, i, \bar{a}', d')$ is a cubic polynomial in the four variables $\bar{g}, i, \bar{a}', d'$:

$$p^d(d/(\bar{g}, i, \bar{a}', d')) = \frac{1}{\sigma(\bar{g}, i, \bar{a}', d')\sqrt{2\pi}} \times \exp\left[-\frac{1}{2}\left(\frac{d-d'}{\sigma(\bar{g}, i, \bar{a}', d')}\right)^2\right]$$

Eq. 19

Using the protein database to extract values for $\bar{g}, i, \bar{a}', d'$ in the discrete ranges shown in Table 10 allowed the derivation of distribution histograms describing the variation of $d'-d$ for various values of each of the four variables. The coefficients of the polynomial used to describe

$\sigma(\bar{g}, i, \bar{a}', d')$ were obtained by least squares fitting to the distribution histograms. The final form obtained is:

$$\begin{aligned} \sigma(g, i, a', d') = & 0.849 - 2.033g - 1.227i + 0.971a' + 1.467d' + 1.382g + 1.539gi - 0.504ga' - 0.259gd' \\ & + 2.412i^2 - 1.496i\bar{a}' - 3.094id' - 0.425\bar{a}'^2 + 0.670\bar{a}'d' - 0.159d'^2 - 0.307\bar{g}^3 - 0.213\bar{g}^2i \\ & + 0.088\bar{g}^2\bar{a}' + 0.020\bar{g}^2d' - 0.969\bar{g}i^2 + 0.453\bar{g}i\bar{a}' + 0.177\bar{g}id' - 0.058\bar{g}\bar{a}'^2 - 0.042\bar{g}\bar{a}'d' \\ & + 0.020\bar{g}d'^2 - 0.847i^3 + 0.055i^2\bar{a}' + 1.546i^2d' + 0.527i\bar{a}'^2 - 0.220i\bar{a}'d' + 0.066\bar{a}'^3 \\ & + 0.153\bar{a}'^2d' - 0.153\bar{a}'d'^2 - 0.0019d'^3 \end{aligned}$$

Eq. 20

Table 10. Derivation of basis PDF relating model protein to one reference protein derived using single homologous structure model

Feature	Symbol	Start	End	Interval	No. of intervals
Average distance from a gap	$\bar{g}$	0	20	^a	6
Fractional sequence identity	i	0	1	0.2	5
Average residue accessibility (%)	$\bar{a}'$	0	100	20	5
Distance (Å)	d', h'	5	30	5	5
Distance difference (Å)	$d-d', h-h'$	-7.0	7.0	0.5	28

^aThe intervals for $\bar{g}$ are 0 to 2, 3, 4 to 5, 6 to 9, 10 to 16, 17 to 24.

Ranges used for the tabulation of the distributions $W(d-d'/\bar{g}, i, \bar{a}', d')$ and $W(h-h'/\bar{g}, i, \bar{a}', h')$ are shown.

Examples of these Cα-Cα distance difference histograms and PDFs are shown in Figure 1.

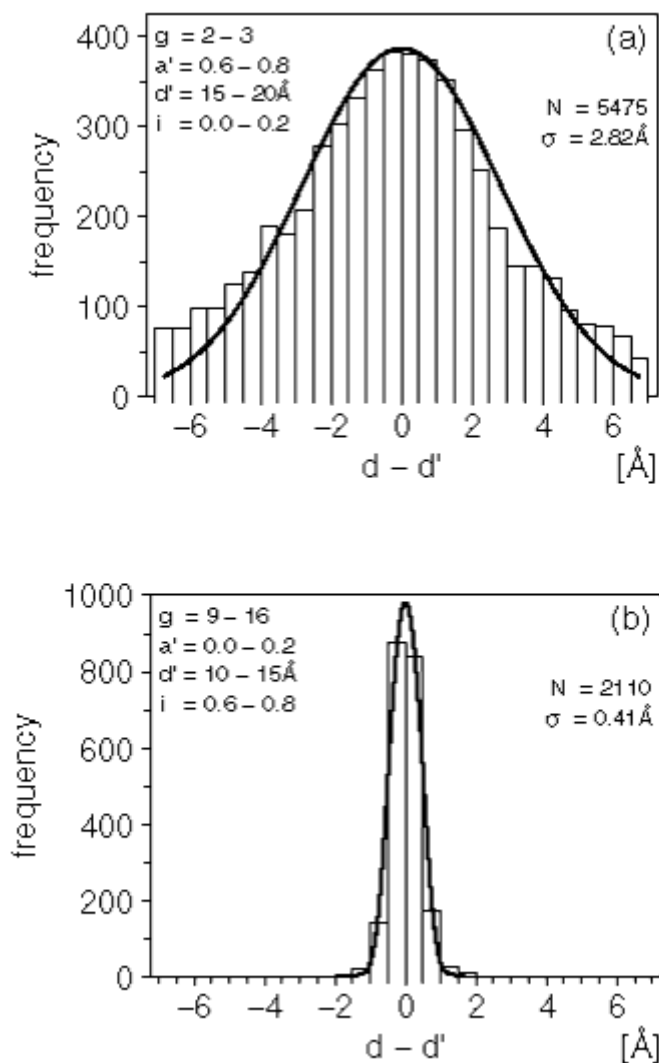


Figure 1. Distribution of the differences between 2 equivalent distances

The histograms show the frequency of the difference between 2 equivalent distances as observed in the alignments database. The curves are least-squares fits obtained as described in the text.

using
modeller

Distance between main-chain N and O atoms

Another main-chain distance constraint is set for N-O atom pairs. The N-N and O-O distances are not constrained. The N-O distance (h) in the model protein is again described by a Gaussian distribution describing the difference between h and the corresponding distance h' in a known structure.

$$p^h(h/\bar{g}, i, \bar{a}', h') = \frac{1}{\sigma(\bar{g}, i, \bar{a}', h')\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{h-h'}{\sigma(\bar{g}, i, \bar{a}', h')}\right)^2\right] \quad \text{Eq. 21}$$

The expression for $\sigma(\bar{g}, i, \bar{a}', h')$ is again a cubic polynomial in these four variables. The coefficients were derived in a manner similar to that just described for the C α -C α distances. The final form obtained for $\sigma(\bar{g}, i, \bar{a}', h')$ is

$$\begin{aligned} \sigma(\bar{g}, i, \bar{a}', h') = & 0.957 - 2.044\bar{g} - 1.078i + 0.995\bar{a}' + 1.477h' + 1.572\bar{g}^2 + 1.148\bar{g}i - 0.525\bar{g}\bar{a}' - 0.483\bar{g}h' \\ & + 1.505i^2 - 0.655i\bar{a}' - 2.849ih' - 0.625\bar{a}'^2 + 0.499\bar{a}'h' - 0.126h'^2 - 0.369\bar{g}^3 - 0.243\bar{g}^2i \\ & + 0.121\bar{g}^2\bar{a}' + 0.067\bar{g}^2h' - 0.592\bar{g}i^2 + 0.346\bar{g}i\bar{a}' + 0.276\bar{g}ih' - 0.032\bar{g}\bar{a}'^2 - 0.061\bar{g}\bar{a}'h' \\ & + 0.036\bar{g}h'^2 + -0.329i^3 - 0.318i^2\bar{a}' + 1.472i^2h' + 0.284i\bar{a}'^2 - 0.293i\bar{a}'h' + 0.382\bar{a}'^3 \\ & + 0.110\bar{a}'^2h' - 0.079\bar{a}'h'^2 - 0.0095h'^3 \end{aligned} \quad \text{Eq. 22}$$

Distances between sidechain-sidechain and sidechain-mainchain atoms

If the model protein contains sidechain-sidechain or sidechain-mainchain atom-atom distances for which an equivalent distance can be found in at least one of the reference proteins, MODELER applies a restraint to that distance in the model. Once again, these are described by a Gaussian distribution of the distance difference, similar to that used above for the main-chain C α -C α and N-O atom-atom distances.

Main-chain conformation

The main-chain conformation classes for a residue can be described by the main-chain dihedral angles (ϕ, ψ). As previously discussed, the third main-chain dihedral angle ω is separately described by a bimodal Gaussian PDF for Pro residues and a monomodal Gaussian PDF for all others. The distribution of ϕ, ψ pairs in the Brookhaven PDB is approximated by six distinct

Gaussian peaks (Sali and Blundell 1993a) whose positions and widths are given in the table below.

Table 11. Parameters of the main-chain conformation classes

	mean		standard deviation	
	ϕ_i	ψ_i	$\sigma_i(\phi)$	$\sigma_i(\psi)$
A	-65	-41	15	15
B	-130	135	15	20
P	-65	140	15	15
G	60	40	10	10
L	90	-10	15	10
E	130	180	25	25

These peaks correspond to six main-chain conformation classes: alpha helix (A), idealized beta strand (B), polyproline conformation (P), the positive ϕ region accessible to Gly residues (G), left-handed alpha helix (L) and an extended conformation (E). If the probability ω_i that a restrained residue in the model is found in the main-chain conformation class i ($i=A,B,P,G,L,E$) is known, then the PDF restraining the dihedral angles ϕ and ψ is given by a binomial function of the form:

$$p^m(\phi, \psi) = \sum_{i=1}^n \omega_i p_i = \frac{1}{2\pi \sigma_{\phi,i} \sigma_{\psi,i} \sqrt{(1 - \rho_i^2)}} \cdot \exp \left\{ -\frac{1}{(1 - \rho_i^2)} \left[\frac{1 - \cos(\phi - \bar{\phi}_i)}{\sigma_{\phi,i}^2} - \rho_i \frac{\sin(\phi - \bar{\phi}_i)}{\sigma_{\phi,i}} \frac{\sin(\psi - \bar{\psi}_i)}{\sigma_{\psi,i}} + \frac{1 - \cos(\psi - \bar{\psi}_i)}{\sigma_{\psi,i}^2} \right] \right\}$$

Eq. 23

where $p < 1$. p is the correlation coefficient between ϕ and ψ . Derivation of the probabilities ω_i used in the above expression depends upon which other features of the protein structure were found to correlate with the main-chain conformation class M . A careful study (Sali and Blundell 1993a) using the local protein database revealed that PDFs of the form $p(M/R, M', s)$ were best suited for predicting ω_i . The variables involved are

R Residue type of the restrained residue.

M' Main-chain conformation class for equivalent residue in reference protein with known structure.

s Residue neighborhood difference.

The residue neighborhood difference s is a value (typical range 0 to 2) that describes the dissimilarity between the residue-residue neighborhoods of two equivalent residues in an alignment. It is computed by the following scheme.

1. Find all neighbors of residue in structure A (cutoff distance 6.0 Å used).
2. Find equivalent residues in model (structure B) using the given sequence alignment.
3. Find sum of dissimilarity scores using a published matrix (Sali and Blundell 1990).
4. Add a gap penalty if there is a deletion in structure B.

The residue neighborhood difference(s) is then the total score divided by the number of neighbors found in structure A. If no neighbors were found in A, then this implies a score of zero. MODELER then uses the conditional PDF $p(M/R, M', s)$ to compute the probabilities ω_i needed to restrain the main-chain conformation class M via Eq. 23. Tests using known structures not involved in the fitting found this scheme to predict the main-chain conformation class with a success rate of approximately 73%. Allowing for swapping between the structurally similar B,P classes as well as between the L,G classes a prediction success of >87% was found (Sali and Blundell 1993a).

Side-chain conformation restraints

In a manner similar to that described above, the side-chain dihedral angles χ_i are restrained using a PDF that is a sum of Gaussians.

$$p^s(\chi_i) = \sum_j \omega_{ij} N[\bar{\chi}_{ij}, \sigma_j(\chi_i)]$$

Eq. 24

where ω_{ij} is the probability that the restrained side-chain angle χ_i is in the class j and $N[\alpha, \sigma]$ is a Gaussian PDF with mean α and standard deviation σ . The dihedral angle ranges defining each class along with the class name, mean, and standard deviations are described in Table 12.

As the table shows, the side-chain conformation classes c_i are dependent upon the residue type r . Thus, the simplest available PDF would be one of

the form $p(c_i/r)$, which is equivalent to using a side-chain rotamer library (Ponder and Richards 1987).

Restraining χ_1 side-chain dihedral angles

All residue types except Val and Ser have χ_1 dihedral angles that most frequently fall into the “-” class. Val predominantly occupies the “t” class, while Ser is slightly more frequently found in the “+” class. A careful study of the effects of including various protein descriptors on the accuracy of predicting the χ_1 side-chain dihedral conformation class c_1 has been performed (Sali and Blundell 1993a). Using the alignments database the optimum PDF for the prediction of the χ_1 class probability ω_{1i} , was found to be of the form $p(\chi_1/r, c'_1, r', s)$. Where r and c_1 are the residue type and χ_1 class of the restrained residue in the model protein, r' and c'_1 are the residue type and χ_1 class of an equivalent residue in a known reference protein, and s is the residue neighborhood difference described previously. This scheme was found to be most accurate in predicting the χ_1 conformation class of large buried residues such as Trp, Cys, Leu, Val, and Tyr where a prediction success of approximately 80% was found. It was least successful in predicting the χ_1 conformation class of smaller exposed residues such as Asn, Met, Arg, Glu, and Ser (approximately 50% successful).

Restraining χ_2 dihedral angles

Residues having a trimodal χ_2 distribution (Asp, Ile, Lys, Leu, Met, Gln, and Arg) prefer the “t” class. Trp is most frequently found in class 1 while His and Asn are most frequently found in class 2. In attempting to predict the χ_2 dihedral angle class c_2 for a given residue in the protein model, the simplest PDF type $p(c_2/r)$ was found to have a 70.7% success rate. Taking into account the residue type r' and the χ_1 and χ_2 dihedral angle classes c'_1 and c'_2 found in an equivalent PDF leads to a PDF of the form $p(c_2/r, r', c'_1, c'_2)$ with a slightly improved prediction success rate of 72.3%. The latter has been used for computing the probabilities ω_{2j} required to restrain χ_2 dihedral angles.

Restraining χ_3 dihedral angles

Glu residues occupy two χ_3 classes, of which class 2 is predominant. The residues Met, Cys, Arg, and Gln occupy three χ_3 classes; Met and Gln have near equal populations for the three classes: Arg occupies the t class in 47% of cases, while Lys does so in 71% of cases. The simplest PDF form $p(C3/r)$ has a χ_3 class prediction ($c3$) success rate of 58.5%. A slight improvement was found when using a PDF of the form $p(c3/r, c3', r', t)$ where $c3'$ is

the χ_3 class of an equivalent residue and t is the main-chain secondary structure class of the model protein. The latter PDF form is used in computing χ_3 class probabilities ω_3 , for use in restraining these dihedral angles via Eq. 24. The prediction success rates based on these PDFs alone are shown below.

Table 12. Definition of χ_1 , χ_2 , χ_3 , and χ_4 side-chain dihedral angle classes

χ type	Residue types	Class	Range(°)	Mean $\bar{\chi}_i$ (°)	Standard deviation (σ)(°)
χ_1	C, D, E, F, H, I, K, L, M, N, Q, R, S, T, V, W, Y	+	[0, 120]	63	10
		t	[120, 240]	180	10
		-	[-120, 0]	-63	10
χ_2	E, I, K, L, M, Q, R	+	[0, 120]	65	10
		t	[120, 240]	180	10
		-	[-120, 0]	-65	10
	D	1	[0, 180]	0	25
	N	1	[-180, 0]	-50	10
		2	[0, 80]	10	10
		3	[80, 180]	140	10
	H, W	1	[-180, 0]	-75	10
		2	[0, 180]	75	10
	F, Y	1	[0, 180]	75	10
χ_3	K, M, R, Q	+	[0, 120]	65	10
		t	[120, 240]	180	10
		-	[-120, 0]	-65	10
	E	1	[35, 85]	60	15
		2	[85, 395]	180	35
χ_4	K	+	[0, 120]	65	15
		t	[120, 240]	180	15
		-	[-120, 0]	-65	15
	R	+	[0, 70]	45	10
		t	[70, 255]	170	35
		-	[-105, 0]	-80	10

Restraining χ_4 dihedral angles

The only standard residues having a χ_4 side-chain are Arg and Lys. These were found to occupy the t class in 64% and 74% of cases respectively. The

prediction success rate of the simplest PDF $p(c_4/r)$ is 75.3%. This form is used for computing probabilities ω_{4j} for restraining χ_4 dihedral angles using Eq. 24.

Basis and feature probability density functions

A feature f of a protein structure is any geometric descriptor associated with a set of atoms $i,j,k,l,\dots$; examples of protein features are a distance between two atoms i,j and a bond angle formed by three atoms i,j,k . The probability density functions (PDFs) described in previous sections are basis PDFs $p^f(f)$ and are derived from a single homologous structure. In general, a single feature in a model protein structure can be restrained by several basis PDFs $p_k^f(f)$ $k=1,2,3,\dots$, where k is the index of the homologous structures. A feature PDF $p^F(f)$ then combines together all appropriate basis PDFs that are needed to describe the feature f . The lower- and uppercase superscripts thus relate to the basis and feature PDFs respectively.

As an example when a certain C α -C α distance d in a model protein sequence is to be restrained using two homologous reference proteins then there are two basis PDFs. If the equivalent C α -C α distances in the reference proteins are d' and d'' , then the basis PDFs are of the form $p(d/d')$ and $p(d/d'')$. These can be combined as a weighted sum to form a PDF that is conditional on both of the reference distances d' and d'' .

$$p(d/d', d'', \bar{s}', \bar{s}'') = \omega(\bar{s}') \cdot p(d/d') + \omega(\bar{s}'') \cdot p(d/d'') \quad \text{Eq. 25}$$

The weighting factors $\omega(\bar{s}')$ and $\omega(\bar{s}'')$ are functions of the average residue neighborhood differences $\bar{s}'$ and $\bar{s}''$ between the model protein and each of the reference proteins. The functional form of the weighting factors $\omega(s)$ is:

$$\omega(s) = \frac{\omega'(s)}{\sum_j \omega'(s_j)}$$

where:

$$\omega(s) = a + \exp(bs^c): \quad \sum \omega(s_j) = 1$$

Eq. 26

The sum j runs over all reference proteins and a, b, c are constants found by a least squares analysis using all possible alignments of three proteins in the

alignments database. The effect of these weighting functions is that the contribution that a related structure makes to a 3D model falls faster than linearly with respect to the average residue neighborhood difference between the modeled and reference protein. This provides the mechanism by which the local sequence similarity between the reference and the model proteins may be applied differently in different regions of the model fold. Since the atoms involved in the $C\alpha$ - $C\alpha$ distance must also obey the van der Waals restraints, the $C\alpha$ - $C\alpha$ conditional PDF is multiplied by a van der Waals basis PDF to yield the final feature PDF.

$$p^D(d) = [\omega_1 p_1^d(d) + \omega_2 p_2^d(d)] p^v(d)$$

Eq. 27

This method of taking a weighted sum is then used for combining any number of basis PDFs of the same type. When generating feature PDFs for features such as main-chain and side-chain conformation, the average residue neighborhood difference $\bar{s}$ is replaced by the residue neighborhood difference s .

A description of the form of the feature PDFs is now given, where the right hand side of each expression contains the basis PDFs that have already been discussed. The variables $a, b, c, \dots$ are the features of the reference structures that are used in deriving the appropriate basis PDFs; the subscript i refers to each reference protein contained in the alignment with the sequence of the model protein, and the weights ω_i are derived from Eq. 26.

Feature PDFs used for restraining model protein features

Ca-Ca distance feature PDF (page 166)
 Main-chain N-O distance feature PDF (page 167)
 Stereochemical feature PDF (page 167)
 Main-chain conformation feature PDFs (page 167)
 Side-chain dihedral angle feature PDFs (page 168)
 Molecular PDF used for structure generation (page 168)

$C\alpha$ - $C\alpha$ distance feature PDF

The feature PDF used to restrain $C\alpha$ - $C\alpha$ distances (d) in the model protein is of the form:

$$p^D(d) = p^v(d) \sum_i \omega_i p_i^d(d/a, b, \dots)$$

Eq. 28

This PDF is computed for all C α -C α atom pairs in the model protein that have at least one equivalent distance in the known reference proteins less than a cutoff distance (20 Å) and where the two residues in the model are not adjacent in the sequence. The summation is over all reference structures having an equivalent C α -C α distance.

Main-chain N-O distance feature PDF

A distance (h) between main-chain N and O atoms is restrained by a feature PDF of the form:

$$p^H(h) = p^v(h) \sum_i \omega_i p_i^h(h/a, b, \dots)$$

Eq. 29

This is evaluated for all N-O distances in the protein model that have at least one equivalent distance in the reference proteins less than a cutoff (10 Å) and where there are at least two intervening residues. As above the sum runs over all known structures having an equivalent N-O pair present.

Stereochemical feature PDF

The PDF form used to restrain a stereochemical feature (e) such as a bond length, bond angle, dihedral angle, or van der Waals contact is simply equal to the basis PDF:

$$p^E(e) = p^e(e)$$

Eq. 30

The feature PDF $p^V(v)$ restrains only those pairs of atoms not already restrained by a bond length, bond angle, or dihedral feature PDF.

Main-chain conformation feature PDFs

The main-chain conformation in the model protein is restrained by a feature PDF of the form:

$$p^M(\phi, \psi) = \begin{cases} \sum_{i=1}^n \omega_i p_i^m((\phi, \psi)/a, b, \dots) & n > 0 \\ p^m((\phi, \psi)/R) & n = 0 \end{cases} \quad \text{Eq. 31}$$

If there is no equivalent residue in any of the known structures ($n = 0$) then the PDF depends only on the residue type. If there are equivalent residues ($n > 0$) then the feature PDF is a weighted sum of basis PDFs.

Side-chain dihedral angle feature PDFs

The side-chain dihedral angles χ_1 , χ_2 , χ_3 , and χ_4 are restrained by feature PDFs of the form:

$$p^S(c) = \begin{cases} \sum_{i=1}^n \omega_i p_i^s(c/a, b, \dots) & n > 0 \\ p^s(c/R) & n = 0 \end{cases} \quad \text{Eq. 32}$$

where c stands for either χ_1 , χ_2 , χ_3 , or χ_4 . In the case of no equivalent residues in the known structures ($n = 0$), a rotamer library based on the residue type is used; when equivalent residues are present ($n > 0$), then the feature PDF is again a weighted sum of basis PDFs.

Molecular PDF used for structure generation

The objective function used for structure generation in MODELER is referred to as a molecular PDF. This is a combination of all the feature PDFs $p^F(f_i)$ used to restrain particular geometric features (f_i) of the protein model. We seek the 3D protein model that is consistent with the most probable values of the individual features. The feature PDFs measure the probability of occurrence for the values of a single feature. The molecular PDF then measures the probability of occurrence for values adopted by several features simultaneously.

If the feature PDFs are assumed to be independent, then the molecular PDF is simply a product of feature PDFs described above:

$$P = \prod_i p^F(f_i)$$

Eq. 33

Finding a maximum for P leads to the most probable 3D model given the alignment of the sequence of the model with those of the known proteins. The assumption of the independence of the feature PDFs is equivalent to describing the system by a molecular energy function that is a sum of terms due to each feature PDF.

Structure generation using MODELER

The previous sections describe the procedures used to derive basis PDFs from a local protein database and to compute the feature PDFs and molecular PDF for a model protein structure. The objective function (F) that is optimized by MODELER is the natural logarithm of the molecular PDF.

$$F = -\ln(P)$$

Eq. 34

Minimizing F corresponds to maximizing P but has advantages in terms of avoiding floating-point overflow issues. The optimization takes place in two distinct stages. The first is a Cartesian space variable target function (VTF) conjugate gradient optimization. The optimum of the MODELER PDF is sought by a series of target functions that eventually equals the full molecular PDF. The series starts only with sequentially local restraints, adding longer range restraints as the calculation proceeds, until all restraints are being used and the full molecular PDF is calculated. The optional second optimization stage consists of a restrained molecular dynamics (MD) simulated annealing scheme. See [Figure 2](#) below.

using
modeler

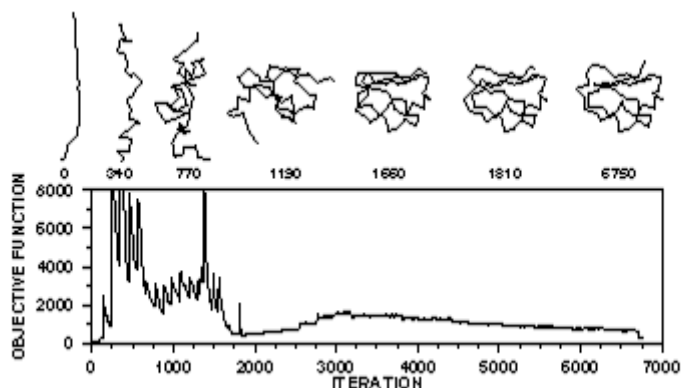


Figure 2. Optimization of the objective function for a linear starting model

Optimization of the objective function (curve) starts with a random or distorted template structure. The iteration number is indicated below each sample structure. The first ~2,000 iterations correspond to the variable target function method relying on the conjugate gradients technique. This approach first satisfies sequentially local restraints and slowly introduces longer range restraints until the complete objective function is optimized. In the last 4,750 iterations, molecular dynamics with simulated annealing is used to refine the model.

The starting conformations used by MODELER are generated by the following scheme.

1. First superpose all templates on the first template using the C α atoms and the given alignment. Each atom in the modeled protein having an equivalent atom in at least one of the templates has coordinates that are the mean of the coordinates for the template atoms. The remaining undefined atoms have their coordinates set by using internal coordinates derived from a CHARMM residue topology library.
2. Next, a random coordinate shift (default between ± 4 Å) is added to each atomic position. This scheme generates starting models similar to the templates for faster convergence.

Modeling loops

For the parts of the model sequence which are not aligned to any template structures, no homology restraints, such as C α -C α distance restraints, can be applied. During model generation, those parts of the model structure are optimized according to CHARMM-derived stereochemical and non-bonded restraints, as well as statistical preferences of the different residue types for the different regions of the Ramachandran plot and for the different sidechain rotamers. Those parts of the structure generally have larger errors compared to the regions which are modeled based on a template

structure. Sometimes, it is desirable to refine the loop conformation after model generation using a statistical pair potential by a more extensive molecular dynamics/simulated annealing procedure. This section discusses the energy function used by loop optimization.

Loops can be defined automatically from the model to template sequence alignment. Any part of the model sequence not aligned with a template structure is defined in MODELER as a loop. You can also specify loops as any sequence segment. During loop optimization, all the atoms are fixed except those defined in the loop region.

The energy of the loop:

$$E = \sum_{\substack{\text{atoms } q \text{ and/or } r \\ \text{in loops}}} E_{k_{seq}(q,r)}^{a_q a_r}(l)$$

Eq. 35

is a sum of the pairwise interactions between loop atoms and the rest of the protein. Those interactions are calculated using the statistical energy derived from pairwise atomic interactions in known protein structures using the Boltzmann approximation (Melo and Feytmans 1997). The energy term is dependent on the stereochemical type of atoms. a_q and a_r represent the number of peptide bonds (sequence separation), $k_{seq}(q,r)$, and class l of the distance between them.

Each pairwise potential $E_k^{ij}(l)$ is derived from the statistics of atom type pairs in known protein structures:

$$E_k^{ij}(l) = RT \ln(1 + M_{ijk} \sigma) - RT \ln \left(1 + M_{ijk} \sigma \frac{f_k^{ij}(l)}{f_k^{xx}(l)} \right)$$

Eq. 36

where M_{ijk} is the number of observations for the atom-type pair ij at the sequence separation k corresponding to:

$$M_{ijk} = \sum_{i=1}^n f(i, j, k, l)$$

Eq. 37

σ ($\approx 1/50$) represents the weight given to each observation.

$f_k^{ij}(l)$ is the relative frequency of occurrence for the atomic pair ij at sequence separation k in the class of distance l :

$$f_k^{ij}(l) = \frac{f(i, j, k, l)}{M_{ijk}}$$

Eq. 38

$f_k^{xx}(l)$ is the relative frequency of occurrence for all the atomic pairs at sequence separation k in the class of distance l :

$$f_k^{xx}(l) = \frac{\sum_{l=1}^n \sum_{j=1}^n f(i, j, k, l)}{\sum_{l=1}^n \sum_{i=1}^n \sum_{j=1}^n f(i, j, k, l)}$$

Eq. 39

The same method and energy function are used to optimize a mutant structure. One or more residues can be selected for mutations from a known protein structure and the structure of the mutant is optimized using the loop optimization protocol. All atoms within a defined distance to the mutated residue are optimized.

References

- Bernstein, F.C.; T.F.Koetzle, G.J.B.Williams, E.F.Meyer Jr, M.D.Brice, J.R.Rodgers, O.Kennard, T.Schimanouchi, M.Tasumi, "The protein data bank: A computer based archival file for macromolecular structures," *J.Mol. Biol.* **112** 535-542 (1977).
- Blundell, T. L.; B. L. Sibanda; M. J. E. Sternberg; J. M. Thornton "Knowledge-based prediction of protein structures and the design of novel molecules" *Nature* (London) **326** 347-352 (1987).
- Blundell, T. L.; D. Carney; S. Gardner; F. Hayes; B. Howlin; T. Hubbard; J. Overinton; D. A. Singh; B. L. Sibanda; M. Sutcliffe "Knowledge-based protein modeling and design" *Eur. J. Biochem.* **172** 513 (1988).
- Browne, W. J.; A.C.T. North; D.C. Phillips; K. Brew; T. C. Vanaman; R.C. Hill "A possible three-dimensional structure of bovine -lactalbumin based on that of hen's egg-white lysozyme" *J. Mol. Biol.* **42** 65-86 (1969).
- Braun, W. and N. Go "Calculation of protein conformations by proton-proton distance constraints: A new efficient algorithm" *J. Mol. Biol.* **186** 611-626 (1985).

- Brooks, B.R.; R.E. Brucoleri, B.D. Olafson, D.J. States, S. Swaminathan, M.Karplus, *J.Comp.Chem.* **4** 187 (1983)
- Greer, J. "Model for haptoglobin heavy chain based upon structural homology" *Proc. Nat. Acad. Sc. U.S.A.* **77** 3393 (1980).
- Greer, J. "Comparative model-building of the mammalian serine proteases" *J. Mol. Biol.* **153** 1027-1042 (1981).
- Greer, J. "Model structure for the Inflammatory Protein C5a" *Science* **228** 1055 (1985).
- Matsumoto, R., A. Sali, N. Ghildyal, M. Karplus, R. L. Stevens "Packaging of proteases and proteoglycans in the granules of mast cells and other hematopoietic cells" *J. Biol. Chem.* **270** 19524- 19531 (1995).
- Melo, F. & Feytmans, E. "Novel knowledge-based mean force potential at the atomic level" *J. Mol. Biol.* **267** 207-222 (1997).
- Nilsson, L.; M. Karplus, *J. Comp. Chem.* **7** 591 (1986).
- Ponder, Richards, *J.Mol. Biol.* **194** 775-791 (1987).
- Sali, A. "Modeling mutations and homologous proteins," *Curr. Opin. Biotech.* **6** 437-451 (1995a).
- Sali, A. "Protein modeling by satisfaction of spatial restraints" *Molecular Medicine Today* **1** 270-277 (1995b).
- Sali, A.; T.L. Blundell, "Definition of general topological equivalence in protein structures: A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming," *J. Mol. Biol.* **212** 403-428 (1990).
- Sali, A.; T.L. Blundell, "Comparative protein modeling by satisfaction of spatial restraints," *Mol. Biol.* **234** 779-815 (1993a).
- Sali, A.; R. Matsumoto, H.P. McNeil, M. Karplus, R.L. Stevens, "Three-dimensional models of four mouse mast cell chymases. Identification of proteoglycan-binding regions and protease-specific antigenic epitopes," *J. Biol. Chem.* **268** 9023-9034 (1993b).
- Sali, A.; J.P. Overington, "Derivation of rules for comparative protein modeling from a database of protein structure alignments," *Protein Sci.* **31** 582-1596 (1994).
- Sali, A.; L. Pottertone, F. Yuan, H. van Vlijmen and M. Karplus "Evaluation of comparative protein modeling by MODELLER" *Proteins* **23** 318-326 (1995).
- Shotton, D. M.; H. C. Watson "Three-dimensional Structure of Tosyl-elastase" *Nature* **225** 811 (1970).

A. Conversion of External Sequence Data Files to QUANTA Format

Data generated outside QUANTA can be read into QUANTA and displayed in the Sequence Viewer as graphs or by coloring sequences according to the input data. The import utility **Read Sequence Data File** is a pullright from **Sequences** on the **Files** pulldown. This utility will currently only read data which is in a QUANTA sequence data format which is described below. A demonstration of reading and displaying data from PHD, the EMBL secondary structure prediction server, is described in Chapter 2..

A demo jiffy program which converts data from the PHD format to QUANTA format can be built from two source files:

```
$QNT_ROOT/user_group_files/sequence_data/phd_quanta.f  
$QNT_ROOT/user_group_files/sequence_data/sequparam.inc
```

To build the program, copy these files to your area and enter the following command:

```
> f77 -o phd_quanta phd_quanta.f
```

The resultant executable program, `phd_quanta`, should read the demonstration PHD output file:

```
$QNT_ROOT/user_group_files/sequence_data/dfr.phd
```

and generate a QUANTA data file identical to:

```
$QNT_ROOT/user_group_files/sequence_data/dfr.sqdat.
```

Two subroutines in the demo program may be useful for anyone writing their own conversion jiffy: *writetitle* which writes the title line for the file *writerealdata*, which writes out one dataset in the appropriate format

Sequence Data File Format

QUANTA reads the ascii file in free format so the exact formatting of each line is flexible. The binary file is defined the same as the ascii file but without the formatting.

The file has two header lines which are defined in FORTRAN:

```

character*30 title      ! title for data file
integer n_data_sets    ! number of datasets in file
integer max_data_length ! maximum number of elements in a
dataset

```

```

write(*,fmt='(a30)')title
write(*,fmt='(i7,1x,i7)')n_data_sets,max_data_length

```

and then for each of the n_data_sets data sets:

```

        character*6 type      ! data type (current only
'REAL' supported)
        character*30 label    ! a label for the dataset
        character*30 seq_name ! the name of the sequence
to which data should
c                                be attached (can be left
blank)
        integer color        ! the color for graph dis-
play
        integer visibility   ! if >0 then graph should be
visible by default
        integer data_length  ! number of elements in data
list
        real data(max_data_length)! the data
c
        type='*REAL'
c
        write(*,fmt='(a6,a30)')type,label
        write(*,fmt='(a30,1x,2(i4,1x),i7)')seq_
name,color,appear,data_length
        write(*,fmt='(10f8.3)')(data(i),i=1,data_length)

```

Sequence User Color File Format

When a user data file is read into QUANTA and you opt to color sequences according to the data in that file then the file seq_user_color.dat is read. QUANTA searches for this file in the standard search path (the current user's working directory, your library directory (defined by environment variable \$QNT_USR), or the QUANTA library directory (defined by environment variable \$HYD_LIB)).

Where the data from the sequence data file is used to color a sequence it is necessary to define the relationship between the data value and the color that the residues are drawn. A simple example might be that all residue with data values less than zero are colored blue and all residues with data values above zero are colored red. Each color scheme in the user color file has a label which should match the label(s) of dataset(s) in the user data file.

The demonstration user color file (\$QNT_ROOT/user_group_files/sequence_data/seq_user_color.dat) defines two coloring schemes for SECTSTR, the secondary structure, and ACCESS, the predicted residue solvent accessibility.

For each coloring scheme there is a label line which begins with an asterisk (*) and can be defined as follows:

```

        character*30 label      ! color scheme label
        character*2  operation ! the operation to apply in
coloring
        integer n_colors      ! number of colors in scheme
        real data_
limit(*)      ! the data cutoff value for color
        integer color(*)      ! the color

        write(*,fmt='(a30,1x,a2)')label,operation
        do 10 n=1,n_colors
        write(*,fmt='(f8.3,1x,i3)')data_limit(n),color(n)
10      continue

```

The parameter operation can be either LT or GT, that is “less than” or “greater than”. If, for example the operation is LT, then the color scheme will be interpreted as:

“If the data value is less than data_limit(n) then color the residue color(n).”

B. Creating a Fragment dmfile

The following procedure creates the database used by the Search Fragment Database utility.

1. **Select from the Brookhaven database a set of protein coordinates files, that have good resolution and include different structure types.**
2. **Construct a file, dmfile, containing a list of these protein coordinate files.**

The following format is used in the file construction:

Number of proteins to be used (NPROT)

Name of coordinate file 1

Name of coordinate file 2

.

Name of coordinate file NPROT

3. **Run the program \$HYD_MSF/dmprep.**

The program prompts for the name of the file, dmfile, which contains the list of proteins, and asks for a name for the distance matrix file to be created, dmfile.new. The program then reads each protein coordinate file and constructs the distance matrix file. The program also creates a QUANTA input command file. This file is later used from within QUANTA to generate an MSF for each of the protein coordinate files. You are prompted for a name for this file.

The dmprep executable distributed with QUANTA is dimensioned for 2000 proteins with limits of 2000 residues and 100,000 alpha carbon distances per protein. The FORTRAN sources for dmprep, dmprep.f and dmsubs.f, are also distributed. This allows you the flexibility to increase the dimensions as needed. To create a modified dmprep executable, type:

```
> cp dmfile.new $HYD_DMf
```

4. Move the distance matrix file to the \$QNT_ROOT/dmatrix directory, renaming it to dmfile.

Because the variable \$HYD_DMF is already defined in the QUANTA environment as \$QNT_ROOT/dmatrix/dmfile, this is easily accomplished by typing:

```
> cp dmfile.new $HYD_DMF
```

where:

dmfile.new = the filename of the distance matrix file
created in step 3.

5. To create the MSFs, which are also required, start QUANTA and type @command_file, where command_file is the name given to the QUANTA command file.

Respond appropriately to the dialog boxes. Treat the sixth character in the atom field as a disorder using the no-hydrogen dictionary file, and exclude symmetry in the molecular structure file.

6. Move the newly created MSFs to the directory \$MSF_LIB.

C. Customizing the Databases

Overview

To use the QUANTA databases requires two data files, that by default are in the database directory, `$QNT_ROOT/library`. These files are `database.dat`, which contains structure information, and `pdbsequence.lib`, that contains sequence information. Versions of these files containing the data extracted from all of the current structures in the Brookhaven Protein Data-bank are included in the PDB/Database release tape. To display fragments corresponding to database search hits, QUANTA reads the coordinates from the appropriate MSF file.

Users may wish to customize the databases to include additional structures. Facilities are included to recreate or append to the database files. The database creation program is `$QNT_ROOT/db/crebase`. This program requires a control file, the PDB master file. The standard distribution version of this file is `$QNT_ROOT/db/crebase/crebase.inp`. This file contains a list of PDB files that go into a database. For each PDB file, the PDB master file specifies what structural analysis should be performed and which database files should be included.

The PDB Master File

The PDB master file, `pdb_master.list`, contains a list of every file, available from Accelrys, in the Brookhaven Protein Databank. Additionally some basic information is found, such as molecule type, crystallographic refinement method, commands for the amount of analysis to perform by the CREBASE program, and which information to stored.

The structure and sequence database files are created by extracting the required data from PDB files. The residue IDs and atom coordinates are taken from the PDB file, and the residue geometry, such as torsion angles and side-chain centers, is calculated. The protein resolution and the information in the PDB cards `HEADER`, `COMPND`, and `SOURCE` are incorporated into the database. Some additional textual information is taken from the CREBASE input file; much of this data is present in the PDB file header but not in a form easily extracted automatically. If you wish to incorporate extra text information into the database, this can be included in the CREBASE input file.

	CREBASE reads the PDB master file, pdb_master.list, in free format. Each line should begin with a four letter keyword command:
<i>FILE</i> filename	Name of PDB file. This must have a path name appropriate for the directory from which the CREBASE job is run.
<i>TYPE</i> type	The type of structure may be: - protein - Cα only (protein with only Cα coordinates) - nucleic acid - carbohydrate The default is protein. Structure analysis for database structure search is only possible on protein.
<i>FAML</i> family	The name(s) of protein families to which protein belongs.
<i>XTAL</i> space group	Crystal space group
<i>REFI</i> RefFAML	Family intement method.
<i>NMOL</i> nmol	Number of independent molecules in the PDB file.
<i>LIGD</i> residue_id residue_name full_name	The residue_id and residue_name as found in the PDB file and a fuller name.
<i>ANAL</i>	Structural analysis of this protein is required.
<i>WRIT</i>	Write all the information on this protein to the database file.
<i>MSF</i>	This CREBASE command file is also the command file for the MSFGEN program. This is a keyword for the MSFGEN program and is ignored by CREBASE; it indicates that this protein should be included in the MSF library.

Running CREBASE to create the Database Files

It is easiest to run the CREBASE program in the directory containing the PDB files. The CREBASE program starts by entering the command:

```
> $HYD_EXE/crebase
```

The user is prompted for the name of the input file, the structure database file, the sequence file, and a log file.

If the name of an existing database file is given, there is the option to append the file or overwrite it. While it is possible to append an existing database file, it is not possible to edit existing files. In the case where a log file name isn't provided, then the output is sent to the screen. The use of log files are recommended, so any problems in creating the database can be

identified afterwards. This program may take considerable time to run, depending on the platform configuration.

For each protein, the log file lists: the filename, resolution, and keyword text. If any residue contains an unexpected number of atoms, it is reported as is quite common for side chains to be undefined or have alternate positions. If the structure is found to contain a high proportion of non-amino residues or the residues contain only C α atoms, then the structure is not included in the database. Once the new versions of the database files are created they should be moved to \$HYD_LIB/database.dat and \$HYD_LIB/pdbsequence.lib.

Creating the MSF Library

In running a standard database search, the SEARCH program writes out a selection file which lists the hit fragments in standard QUANTA selection format. For QUANTA to use the selection file it is necessary to have MSF files for the proteins.

The easiest means to generate MSFs is to run QUANTA with a stream file which controls reading PDB files and creating new MSFs. An appropriate stream file can be generated from the PDB master file by running the jiffy program \$HYD_EXE/msfgen. The keyword MSF following a PDB file name in the PDB master file indicates that PDB file should be included in the stream file.

The MSF files are created in the directory from which the QUANTA job is run. The stream file must have appropriate pathnames for the PDB files to be accessed from that directory.

D. The Geometric Structure Definition File

This file provides the definitions of the common amino acids in terms of their atomic composition, intra-residue bonds and structural geometry. The data is used in:

- the Protein Special bonding algorithm, which finds the bonds on the basis of atom name irrespective of geometry;
- the Protein Editor to generate geometrically ideal (though not necessarily energetically good) structures for mutation and insertion and also for editing terminal groups;
- regularization used in several places in Protein Design and the X-ray module.
- Split and Clean to provide the definition of correct protein structure

Note that whether an amino acid is recognized as such and displayed in the Sequence Viewer depends on the data in \$HYD_LIB/protein_param.dat after the keyword AMINO. The protein_structure.gsd supersedes the data in the template files \$HYD_LIB/tmplatnoh/*.pdb, \$HYD_LIB/tmplatpol/*.pdb and \$HYD_LIB/tmplatall/*.pdb which was used in the Protein Editor utility for mutation and insertion and the list of bonded atoms in \$HYD_LIB/protein_torsion.dat used in the Protein Special bonding algorithm.

The default versions of this file are \$HYD_LIB/protein_structure.gsd and its binary version \$HYD_LIB/protein_structure.bgsd. The binary version of the file is normally read by QUANTA as this speeds up the access considerably.

If you need to edit the geometric structure file, copy the ascii file to your own directory, edit it and then run QUANTA and type in the command line **TMPL**. You will be asked for the name of the ASCII file, which is converted to binary. QUANTA then reads your local copy of protein_structure.bgsd (in either the current working directory or your data directory) in preference to the version in \$HYD_LIB.

The file format

The file uses four-letter keywords to identify data. The data is read free format, so exact formatting is not necessary but the spaces between are essential.

NAME *code full_name*

code : a short name of four or fewer characters

full_name: name of maximum 40 characters

TYPE *type*

type: the functional type of the group.

The recognized types are: PROTEIN_N_TERMINAL, PROTEIN_C_TERMINAL, PROTEIN_MAIN, PROTEIN_SIDE

GROUP *n_atoms*

n_atoms: number of atoms in the group as defined after the following *n_atoms* lines.

*name x y z type_allH charge_allH type_polarH charge_polarH
type_noH charge_noH*

name: maximum four character atom name

x,y,z: coordinates for this atom in a group of ideal geometry.

type_allH, charge_allH: the atom type and charge for an all atom representation of the group

type_polarH, charge_polarH: the atom type and charge for a polar hydrogen only representation of the group

type_noH, charge_noH: the atom type and charge for a no hydrogen representation of the group

Here the atom types use the QUANTA number scheme which corresponds to CHARMM atom types. The atom types are listed in the file \$HYD_LIB/param.par Where a hydrogen atom is excluded from a representation it is given an atom type of zero.

TEMPlate

is followed by three lines of:

name x y z

name atom name of maximum four characters x y z coordinates

The template is the three neighboring atoms to a group which are used, for example in mutating a protein side chain, as a guide for positioning the group.

BOND *name_1 name_2 name_10*

name_1 name_2...: a list of maximum five pairs of atom names which indicate the bonded atoms

IMPRoper *representation name_1 name_2 name_3 name_4*

defines any improper dihedrals

representation: the minimal hydrogen representation for which the improper is valid. For example, an improper involving a polar hydrogen would not be valid in a no-hydrogen representation so the appropriate representation term is "POLAR_H". The recognized terms are: NO_H, POLAR_H, ALL_H.

name_1 ... name_4: names of four atom which should be constrained by an improper dihedral.

DIHEdral *representation name_1 name_2 name_3 name_4 angle*

Identify dihedral torsions in the group representation: minimal hydrogen representation for which dihedral is appropriate - see explanation after the IMPRoper keyword.

name_1 ... name_4: names of four atom which define the dihedral

angle: the optimal value of the dihedral

REType *old_name new_name type_allH charge_allH type_polarH charge_polarH type_noH charge_noH*

Where the addition of a group to a residue causes a change in the atom name, type or charge of another atom in the residue this keyword is used to define the new name, type and charge.

old_name: old atom name

new_name: new atom name

type_allH, charge_allH: the new type and charge for the atom assuming an all atom representation

type_polarH, charge_polarH: the new type and charge for the atom assuming a polar hydrogen representation

type_noH, charge_noH: the new type and charge for the atom assuming a no-hydrogen representation

ALTName *name alt_name*:

where atoms in a residue have alternative names in common use *name* is the atom name used here and *alt_name* is the alternative name.

E. The Protein Parameter File

This file contains a variety of parameters used in protein modeling. Each parameter is identified by a four-letter keyword and the data for that parameter is terminated with the keyword END. Several data types involve two-dimensional matrices of values for one amino acid type against another amino acid type. The order in which the amino acids appear in the matrix is defined by the keyword ORDER followed by a list of the 4-letter codes for the amino acids in the appropriate order. The list is terminated with an asterisk (*).

AMINO acid names

This is a list of the residue names which would be recognized as amino acids and displayed in the Sequence Viewer. Any other residue names in a mainly-protein MSF would be regarded as solvent or ligand.

The format of the list is:

name code substitute

name: four letter residue name

code: single character code used in the **Sequence Viewer**

substitute: for none standard amino acids many parameters have not being derived but parameters for a similar standard amino acid will be substituted if the four-letter name of the similar standard amino acid is entered.

SOLVent

This keyword is followed by a list of four-letter residue names which will be recognized as water. This information is used in the protein specific display utility **Molecule Colors** on the **Protein Utilities** palette.

CLASS

This is followed by definitions of the classifications used in coloring sequences:

NAME *classification*

group_type color name_1 name_2 ... name_n

classification: the name of the classification (e.g. Hydrophobicity) This name will appear in the Molecule Colors dialog box for you to select coloring by that classification.

group_type: the name for the type of a group of amino acids color: the color that will be applied to a the group of amino acids

name_1, name_2, ...name_n: the four-letter names for the amino acids which belong to this group

EQUIvalent torsions lookup

In modeling side chain conformations may be copied between equivalent residues in homologous structures. If the residues are not of identical amino

acid type then copying side chain torsions is only reasonable where the side chains have some structural similarity. The following table gives the maximum number of torsions that are copied between pairwise combinations of amino acid types.

PARAM name

Parameters which have one value for each amino acid type are defined. Currently only the ACCESS parameter, the maximum solvent accessible area, is defined here.

ROTamer rotamer_type

name n_rotamer
sec_str frequency tor_1 tor_2 ... tor_n

rotamer_type: the name of the author of the rotamer library name: four character residue name n_rotamer: the number of rotamers for that residue type - the rotamers are list on the following lines

sec_str: if the rotamer is specific for one secondary structure type then H (helix) or E (extended) are indicated

frequency: the observed percentage frequency of that rotamer tor_1, tor_2 ...

tor_n: the ideal torsion values for the rotamer. These are not necessary for all the torsions in the sidechain.

FOLD

n_res sec_str_code name
phi_1 psi_1
phi_2 psi_2
...
phi_n psi_n

The standard backbone folds are defined here and are used to fold a chain in a standard structure in the **Apply Conformation** utility in the **Model Backbone** application.

n_res: number of residues in the fold. A negative value implies that the fold can be extended indefinitely.

sec_str_code: A numerical code for the secondary structure type used within QUANTA. The values are:

-3, -4	beta bulge
-2	beta strand
-1	possible beta strand with appropriate conformation but without correct hydrogen bonding
0	undefined

1	possible alpha helix with appropriate conformation but without correct hydrogen bonding
3	3-turn
4	4-turn
5	5-turn
6	alpha helix

name: the name of the fold which will be used in the interface dialog box

phi_n, psi_n: The main chain phi and psi angles, in degrees, for *n_res* residues. A value of -999.9 implies that there is no defined value for this torsion.

protein
parameter file

F. Read Sequence File Formats

Overview

The file formats recognized are:

- FASTA (.aa extension)
- GCG (.gcg extension)
- NBRF-PIR (.pir extension)
- Swissprot (.sws extension)

The first three formats are the easiest to use. When necessary, the names and extensions can be changed to something more appropriate. Sequences can be in either uppercase or lowercase. All formats recognize the one-letter code for the 20 standard amino acids plus:

B = ASX (asp or asn)

Z = GLX (glu or gln)

X = UNK (unknown)

Many file formats contain additional comment lines, which are ignored when the files are read.

read seq
file formats

Pearson (FASTA) format (extension .aa)

The first line of the file is the title and begins with a “>”. The rest of the record on the line is the title. The sequence is read until a “*” or end of file is encountered. Spaces and punctuation characters are ignored.

```
>P1REIA1 BENGE-*JONES IMMUNOGLOBUL: 1 A 107 A L=107 NRES= 214
D I Q M T Q S P S S L S A S V G D R V T I T C Q A S Q D I I K Y L N
W Y Q Q T P G K A P K L L I Y E A S N L Q A G V P S R F S G S G S G
T D Y T F T I S S L Q P E D I A T Y Y C Q Q Y Q S L P Y T F G Q G T
K L Q I T *
```

GCG (extension .gcg)

The GCG file may contain an arbitrary number of lines of comment at the start of the file. These are followed by a blank line and then a title line. The sequence is given with 50 residues per line, and each line beginning with

the sequence number. These sequences can be obtained from the GCG package by entering command:

```
>  FETCH  -DOCL= x
```

and then entering the appropriate code for the sequence.

The value for *x* represents the number of documentation lines. For example to obtain a haemoglobin sequence the following command and code is used;

```
>  fetch  -docl = 5
```

HAHU

This gives the sequence with five lines of documentation. There are two blank lines. One occurs after the documentation and the other before the sequence.

```
P1;HAHU - Hemoglobin alpha chain - Human, chimpanzee, and pygmy
chimpanzee
C;Species: Homo sapiens (man); Pan troglodytes (chimpanzee); Pan
paniscus(pygmy chimpanzee, bonobo)
C;Accession: A02248
R;Michelson, A.M., and Orkin, S.H. . . .
HAHU Length: 141 January 26, 1993 11:37 Type: P Check: 9231 ..

1      VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSH
51     GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKL
101    LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTISKYR
```

Using zero lines of documentation the retrieved sequence would appear as:

```
HAHU Length: 141 January 25, 1993 15:21 Type: P Check: 9231 ..

1      VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSH
51     GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKL
101    LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTISKYR
```

The **Read Sequence** facility reads any gcg sequence, provided the records with the sequence information are as shown above. The placement of the integer field showing the sequence numbers is important.

NBRF-PIR

The first line contains a > at position 1 and a; at position 4, followed by the sequence ID. The second line is a title which is followed by the sequence that ends with a *. Spaces are ignored. Several different examples follow of this type of sequence format.

```
>P1;HAHU Hemoglobin alpha chain - Human, chimpanzee, and pygmy
chimpanzee
V L S P A D K T N V K A A W G K V G A H A G E Y G A E A L E R M F L
S F P T T K T Y F P H F D L S H G S A Q V K G H G K K V A D A L T N
A V A H V D D M P N A L S A L S D L H A H K L R V D P V N F K L L S
H C L L V T L A A H L P A E F T P A V H A S L D K F L A S V S T V L
T S K Y R *
```

```
>P1;CHOA$STRSQ NRES 546 (T= 74 ) DE CHOLESTEROL OXIDASE PRECURSOR
(EC 1.1.3.6) (C
```

```
MTAQQHLSRRRLGMAAFGAAALAGGTTIAAPRAAAAKSAADNGGYVPA
VVIGTGYGAASALRLGEAGVQTLMEGQLWNQPGPDGNIFCGMLNPDK
RSSWFKNRTEAPLGSFLWLDVVNRNIDPYAGVLDRVNYDQMSVYVGRGVG
GGSLVNGGMAVEPKRSYFEEILPRVDSSEMYDRYFPRANSMLRVNHIDTK
WFEDTEWYKFARVSREQAGKAGLGTVPNNVYDFGVMQREAAAGEVPKSAL
ATEVIYGNNHGKQSLDKTYLAAALGTGKVTIQLHQVKTIRQTKDGGYAL
TVEQKDTDGKLLATKEISCRYLFLGAGSLGSTELLVRARDTGTLPNLNSE
VGAWGPNGNIMTARANHMMNPTGAHQSSIPALGIDAWDNSDSSVFAEIA
PMPAGLETWVSLYLAIITKNPQRGTFFVYDAATDRAKLNWTRDQNAVAVNA
KALFDRINKANGTIYRYDLFGTLKAFADDFCYHPLGGCVLGKATDDYGR
VAGYKNLYVTDGSLIPGSVGVNPFVTITALAERNVERIIKQDVTAS*
```

The following is an example of a gcg sequence converted to PIR using the utility TOPIR in the GCG package:

```
>P1;HAHU hahu.gcg => HAHU
VLSPADKTNVKAAWGKVGAGHAGEYGAELERMFLSFPTTKTYFPHFDLSH
GSAQVKGHGKKVADALTNAAHVDDMPNALSALSDLHAHKL RVDPNVFKL
LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLT SKYR*
C;P1;HAHU - Hemoglobin alpha chain - Human and chimpanzees
```

SWISSPROT (extension .sws)

The file begins with lines of comment which have two-letter keywords at the start of the line. The sequence is preceded by a line beginning with the keyword SQ and is followed by a line beginning //. Spaces are ignored.

```
ID 104K$THEPA STANDARD; PRT; 924 AA.
AC P15711;
DT 01-APR-1990 (REL. 14, CREATED)
DT 01-APR-1990 (REL. 14, LAST SEQUENCE UPDATE)
DT 01-AUG-1990 (REL. 15, LAST ANNOTATION UPDATE)
DE 104 KD MICRONEME-RHOPTRY ANTIGEN.
OS THEILERIA PARVA.
OC EUKARYOTA; PROTOZOA; APICOMPLEXA; SPOROZOA; COCCIDIA; PIROPLASMI
DA.
RN [1]
RP SEQUENCE FROM N.A.
RC STRAIN=MUGUGA;
RC MEDLINE=90158697;
RA IAMS K.P., YOUNG J.R., NENE V., DESAI J., WEBSTER P.,
RA OLE-MOIYOI O.K., MUSOKE A.J.;
RL MOL. BIOCHEM. PARASITOL. 39:47-60(1990).
CC -!- DEVELOPMENTAL STAGE: SPOROZOIT ANTIGEN.
CC -!- SUBCELLULAR LOCATION: IN MICRONEME/RHOPTRY COMPLEXES.
DR EMBL; M29954; TP104MRA.
KW ANTIGEN; PROLINE-RICH; REPEAT.
FT DOMAIN 1 19 HYDROPHOBIC STRETCH.
FT DOMAIN 905 924 HYDROPHOBIC STRETCH.
SQ SEQUENCE 924 AA; 103625 MW; 4746107 CN;
MKFLILLFNILCLFPVLAADNHGVPQGASGVDPITFDINSNQTGPAFLT
AVEMAGVKYL
QVQHGSNVNIHRLVEGNVVIWENASTPLYTGAIVTNNDPGYMAYVEVLGD
PNLQFFIKSG
```

DAWVTLSEHEYLAQLQEIQAQVHIESVFSLNMAFQLENNKYEVETHAKNG
 ANMVTFIIPRN
 GHICKMVYHKNVRIYKATGNDTVTSVVGFFRGLRLLINVSIDDNGMMS
 NRYFQHVDDK
 YVPISQKNYETGIVKLKDYKHAYHPVDLDIKDIDYTMFHLADATYHEPCF
 KIIPNTGFCI
 TKLFDGDQVLYESFNPLIHCINEVHIYDRNNGSIICLHLNYSPPSYKAYL
 VLKDTGWEAT
 THPLLEEKIEELQDQACELDVNFISDKDLYAALTNADLNYTMVTPRPH
 RDVIRVSDGS
 EVLWYYEGLDNFLVCAWIYVSDGVASLVHLRIKDRIPANNDIYVLKGDLY
 WTRITKIQFT
 QEIKRLVKKSKKKLAPITEEDSDKHDEPPEGPGASGLPPKAPGDKEGSEG
 HKGPSKGSDS
 SKEGKKPGSGKKPGPAREHKPSKIPTLSKKPSGPKDKHPRDPKEPRKSK
 SPRTASPTRR
 PSPKLPQLSKLPKSTSPRSPPPPTRPSSPERPEGTKIIKTSKPPSPKPPF
 DPSFKEKFYD
 DYSKAASRSKETKTTVVLDSEFESILKETLPETPGTPFTTPRPVPPKRPR
 PESPFEPK
 DPDSPSTSPSEFFTPPESTRFRHETPADTLPDVTAEKFKEPDVTAETK
 SPDEAMKRPR
 SPSEYEDTSPGDYPSLPMKRHLRLRLTTMETDPGRMAKDASGKPVK
 LKRSKSFDDL
 TTVELAPEPKASRIVVDEGTEADDEETHPPEERQKTEVRRRRPPKKPSK
 SPRPSKPKK
 KKPDSAYIPSILAILVVSILVIGIL
 //

G. Running the Search Standalone

Overview

The Protein Design application includes an interface to the search program, but input files can be created and/or edited and the search run independent of QUANTA. The format of the input file is described below.

Search Commands

Each line of a search command file consists of a four-letter command keyword that is followed by a list of parameters that are free format, separated by spaces. Many of the parameters need not be specified, but if subsequent parameters in the command line will be specified, then a * character should be entered in the command line to signify a missing parameter. Missing parameters are given default values. In the following documentation, the keywords and their arguments are listed. Arguments for which there are default values and which can be given as * are enclosed in square brackets, ([]).

DBAS database_file

The name of the database file. By default the name is \$HYD_LIB/database.dat.

WILD wildcard_file

The name of the wildcard file. By default the name is \$HYD_LIB/wildcard.dat.

MRNG n1 n2

Defines a range of proteins to be searched. The numbers n1 and n2 refer to the position of the protein in the database file. By default, all proteins in the database are searched.

*MNAM name1 [name2]
[name3]*

Enter a list of proteins to be searched. The names should correspond to the original PDB file names. The names will be converted to uppercase. The PDB file names stored in the database are in uppercase. For a long list of file names, this command can be invoked more than once.

MKEY keyword

Search only those proteins with a text string matching the keyword in their description. Note that the text in the description is all uppercase and the keyword text will be converted to uppercase.

Only one keyword should be entered on each command line. This command can be multiply invoked; any protein that matches any of the given keywords will be searched.

RSLN resol

Search only the molecules with resolution less than or equal to resol.

INFO

For each hit protein, extract the header information for that protein from the database file and write it to the log file. This can be used in conjunction with the commands that specify a protein (MRNG, MNAM, MKEY and RSLN) without any structural template being defined.

NHIT nhit

Stop search once nhit hits have been found. The default is 50.

DTOL distance

The distance tolerance on all inter-fragment inter-atomic distance searches. The default is 1.0 Å.

ATOL angle1 angle2

The angle tolerance on torsion angle searches for phi and psi torsion angles. The defaults are 30.0° and 30.0°.

CTOL angle

The angle tolerance on torsion angle searches for C alpha pseudo torsion angles. The default is 50.0°.

STOL angle

The angle tolerance on torsion angle searches for side chain torsion angles. The default is 30.0°.

TMPL nres template_name

Signifies the start of defining a template for a fragment of nres residues. This card must be followed by nres RESD cards. The template is given a name that is used for reference by the CONS command.

*RESD [restyp] [secstr]
[phi] [psi] [cators]*

Specify each residue in a fragment template.

Restyp is the single character code for the residue type or a number for the wildcard template. The default is wildcard. This residue in the template can be any residue type.

Secstr is a character code to denote the secondary structure type:

- H— folded conformation
- A— alpha helix
- T— a turn
- 3— a 3 residue turn
- 4— a 4 residue turn
- 5— a 5 residue turn
- E— extended chain
- N— N terminal residue

If any of these codes are prefaced by a NOT then the search is for residues *not* of that secondary structure type. The default is wildcard.

Phi, psi specifies main chain torsion angle in degrees.

Cators for residue i is the pseudo torsion between the four consecutive C alpha atoms $C\alpha(i-1) - C\alpha(i) - C\alpha(i+1) - C\alpha(i+2)$.

*SIDE [sidtor1 sidtor2
sidtor3 sidtor4 sidtor5]*

Sidtor is the side chain torsions as defined by IUPAC-IUB and listed in Appendix H. This line should follow the RESD card for the residue to which it applies.

DIST type tarres tardis

The DIST card must follow the RESD card for one of the two residues to which it applies. This card is optional to specify interatomic distances between two residues in the same template. The actual distance will be between the C α atom or the sidechain center dependent on the type parameter which may be:

CACAC α –C α distance
CASAC α –sidechain distance
SICAsidechain–C α distance
SISIsidechain–sidechain distance

The distance from this residue to residue tars residues on in the template should be within the tolerance distance DTOL of tardis.

*CONS contype
tmpnam1:res1
tmpnam2:res2 minlim
maxlim [chntest]*

Defines a constraint between two residues in different templates. The constrained parameter is defined by contype:

CACAC α –C α distance
CASIC α –sidechain distance
SICAsidechain–C α distance
SISIsidechain–sidechain distance

IRNGnumber of residues in sequence between
template residues
XRNGexclude range of number of residues in sequence

The residues between which the constraint holds are given in the format:

template_name:residue_number

Minlim and maxlim are the minimum and maximum allowed values of the parameter specified by contype.

Chntest flags to test if the fragments are in the same protein chain. The segment IDs given in the Protein Data Bank file are taken to indicate the different chains. If chntest is zero, no check is performed; if chntest is 1, the fragments must be in the same chain.

DEBUG

Output some diagnostic information to log file.

Running a Search

The standard command line to run the database search is:

\$HYD_EXE/search job_name

If the jobname is omitted, you are prompted for it. The input command file, `job_name.ddb`, has been described above. The search program creates two output files. The log file (`job_name.log`) contains the following:

- A listing of the input command file
- A list of the tests (such as sequence and secondary structure) that have been performed
- Information on the database file used
- For each hit, the name of the protein (the PDB filename).
- For each residue, the residue ID, residue type, and secondary structure code

If a MSF doesn't exist for this protein structure in the MSF library or your working directory, then there is a warning message to that effect.

The selection file (`job_name.sel`) is in standard QUANTA selection format. This enables the selection of all the hits that occur in structures for which there is an MSF either in the MSF library or your working directory. If the MSF isn't found, then the selection file does not include that structure.

H. Torsion Angles and Centers

Overview

The database file contains the sidechain torsion angles and one coordinate for the center of the sidechain. This center is the center of mass or the center of the reactive group of the sidechain. The actual specification for each amino acid residue is listed in the following table.

Table 13. Side Chain Torsion Angles and Centers

ami- no acid	center of mass	chi1	chi2	chi3	chi4	chi5
PRO	C γ	N-C α -C β -C γ	C α -C β -C γ -C δ			
GLY	CA					
ALA	CB					
VAL	C γ 1/C γ 2	N-C α -C β -C γ 1				
LEU	C δ 1/C δ 2	N-C α -C β -C δ	C α -C β -C γ -C δ 1			
ILE	C δ	N-C α -C β -C γ 1	C α -C β -C γ 1-C δ			
MET	C ϵ	N-C α -C β -C γ	C α -C β -C γ -S δ	C β -C γ -S δ -C ϵ		
PHE	CH	N-C α -C β -C γ	C α -C β -C γ -C δ 1			
TRP	N ϵ 1	N-C α -C β -C γ	C α -C β -C γ -C δ 1			
SER	O γ	N-C α -C β -O γ				
THR	O γ 1	N-C α -C β -O γ 1				
ASN	O δ 1/N δ 2	N-C α -C β -C γ	C α -C β -C γ -O δ 1			
GLN	O ϵ 1/N ϵ 2	N-C α -C β -C γ	C α -C β -C γ -C δ	C β -C γ -C δ -O ϵ 1		
CYS	S γ	N-C α -C β -S γ				
TYR	O ζ	N-C α -C β -C γ	C α -C β -C γ -C δ 1			
ASP	O δ 1/O δ 2	N-C α -C β -C γ	C α -C β -C γ -O δ 1			
GLU	O ϵ 1/O ϵ 2	N-C α -C β -C γ	C α -C β -C γ -C δ	C β -C γ -C δ -O ϵ 1		
LYS	N ζ	N-C α -C β -C γ	C α -C β -C γ -C δ	C β -C γ -C δ -C ϵ	C γ -C δ -C ϵ -N ζ	
ARG	NH1/NH2	N-C α -C β -C γ	C α -C β -C γ -C δ	C β -C γ -C δ -N ϵ	C γ -C δ -N ϵ -C ζ	C δ -N ϵ -C ζ -NH1
HIS	N δ 1/C δ 2	N-C α -C β -C γ	C α -C β -C γ -N δ 1			

torsion angles,
centers

Database file format

This format can be defined by the following FORTRAN specification:

```
parameter ( nzres=2000, nzseg=20)
parameter ( nzlig=200, nzfam=100, nznam=100, nzkey=320)
character*8 protnm ! protein name
integer nlig! number of ligand records
integer nrec! number of analysis records for
! protein
integer nxlmol! number of protein molecules
! defined in PDB file
integer nseg! number of segments
integer nres! number of residues
integer natom! number of protein atoms
integer nhetat! number of non-protein atoms
integer nhoh! number of water atoms
real resol! resolution of crystal structure
character*(nzfam)! protein family
character*(nznam)! name of protein
character*(nzkey) descrip! keyword description
character*(nzlig) ligand! description of ligand
character*1 segid (nzseg)! segment name
integer prsseg (nzseg)! first residue in segment
integer nrsseg (nzseg)! number of residues in segment
character*5 idres (nzres)! residue ids
integer rescode (nzres)! residue type code
integer secstr (nzres)! secondary structure type code
real x (3,nzres)! Ca atom coordinates
real xside (3,nzres)! 'average' side chain coordinate
real cators (nzres)! Ca pseudo-torsions
real phipsi (2,nzres)! phi/psi torsion angles
real sidtor (5,nzres)! side chain torsion angles
c write out the header records describing the protein
  write (file) protnm, nlig, nrec, nxlmol, nseg, nres, natom,
1    nhetat,nhoh
  write(file) resol, family, name, descrip
c write the description of ligands - NB variable number of records
  do 10 n=1, nlig
10   write (file) ligand(n)
c write the structural description NB not necessarily present for
c all entries in the database - see the nrec parameter
  write (file) (segid(k), k=1,nseg), (prsseg(k), k=1,nseg),(nrs-
seg(k), k=1,nseg)
  write (file) (idres(k), k=1,nres)
  write (file) (rescode(k), k=1,nres)
  write (file) (secstr(k), k=1,nres)
  write (file) ((x(j,k), j=1,3), k=1,nres)
  write (file) ((xside(j,k), j=1,3), k=1,nres)
  write (file) (cators(k), k=1,nres)
  write (file) ((phipsi(j,k), j=1,2), k=1,nres)
  write (file) ((sidtor(j,k), j=1,5), k=1,nres)
```

I. Wildcard Residue Type File

The file \$HYD_LIB/wildcard.dat contains the definitions of amino acid wildcard types. The information is in the form of a matrix with each column representing one amino acid type and each row representing one wildcard type. A matrix element is a 1 if the amino acid type is included in a wildcard type, or a 0 if the amino acid type is not included in a wildcard type.

The first line of the file contains a number that is the total of the wildcard type definitions found in the file. These definitions include the number of rows minus the first row in the file, and a list of one letter amino acid codes that define the order of the amino acids in the matrix. The subsequent lines contain the number of the row followed by a series of 0s and 1s. This is the definition matrix. Finally the wildcard type code name, a maximum of six characters, and a short definition of the wildcard type.

By default, QUANTA will take wildcard definitions from any file named wildcard.dat in your current working directory. If such a file does not exist, then QUANTA reads the file in \$HYD_LIB. If you wish to customize this file then copy a version of the file from \$HYD_LIB to your working directory, and make the appropriate changes and/or additions. Ensure that you have the correct number of rows defined at the beginning of the first line. The maximum number of allowed wildcard types is 20.

J. MolScript

MolScript is a program for creating molecular graphics in the form of PostScript plot files. Possible representations are simple wire models, CPK spheres, ball-and-stick models, text labels and Jane Richardson-type schematic drawings of proteins, based on atomic coordinates in various formats. Color, grayscale, shading and depth cueing can be applied to the various graphical objects.

The QUANTA interface to MolScript generates the input script and coordinate files to reproduce as nearly as possible the current representation. Objects rendered include bonds, solid models, hydrogen bonds, distance monitors, labels and atom IDs.

The interface is accessed by using the function **Plot Molecules** → **Generate** on the **File** menu. This displays the usual Setup Hard Copy Plot of Molecules dialog, which now includes an extra radio button, **MolScript Plot**. When this is used, a new dialog box is displayed containing the MolScript interface options.

The following is a brief description of the various options:

Root name for files

The interface will produce two files, *filename.pdb* containing the molecular coordinates, and *filename.in* containing the MolScript commands. Running the program will produce the PostScript file *filename.ps* or *filename.eps*.

Output format

Allows the choice of normal (.ps) or Encapsulated PostScript (.eps).

Maximum bond distance

This is used to control the maximum bond distance allowed in ball-and-stick plots. The default, 2.1 Å is suitable for proteins containing disulfide bridges but no hydrogen atoms; if hydrogen atoms are present a smaller value will be needed to prevent the hydrogen atoms being bonded together. Vector bonds are generated using the **LINE** command so that the exact bonding generated by QUANTA can be replicated regardless of this setting.

Draw thin ribbons as turn or coil

MolScript allows for optional smoothing of thin ribbons. This can make for a clearer diagram, but should be avoided if sidechains are also displayed.

Thin ribbon all one color

This can sometimes make for a clearer diagram.

Change to solid color scheme

Changes the color settings to be optimized for the drawing of solid models. Note — the plot will always have a white background regardless of the QUANTA background color.

Retain input molscript and pdb files

Allows the input files to be retained — expert users may wish to manually adjust the input and rerun molscript. If this option is unchecked, then the input files will be deleted if MolScript is run successfully.

Preview plot

This option causes the resulting PostScript plot to be displayed using the xpsview viewing program. This requires that xpsview can be found in the user's path or that there is an alias pointing to it.

If both solid models and vector bonds are displayed on screen, then both will be included in the MolScript plot. Remember that vector bonds can be rapidly toggled using function key <F10>, if the default key bindings are set up. Otherwise enter the commands **HIDE BOND** and **SHOW BOND**.

When the **OK** button is hit, the input files are created and molscript will be run if it can be found in the user's path or if there is an alias pointing to it.

To obtain MolScript

The MolScript program must be obtained from its author, Dr. Per Kraulis.

Per Kraulis, Ph.D
Pharmacia & Upjohn, Inc.
PPC Sweden Research, N62:5
S-112 87 Stockholm
SWEDEN

phone: +46 (8) 695 78 34
fax: +46 (8) 695 40 82
e-mail: krp@sgikrpe.sto.se.pnu.com
or p.j.kraulis@bioc.cam.ac.uk
or molscript@bioc.cam.ac.uk

The software is provided under license — details from Per Kraulis.
In publications, you must refer to:

Per J. Kraulis "MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures" *Journal of Applied Crystallography* **24** 946-950 (1991).

Index

A

accessibility calculations

 solvent sphere 80

active range 33

active sequence 33

Align and Superpose 7

 Constraint Palette 45

 Add Constraint 45

 Delete All 46

 Delete Constraint 46

 Exit Constraints 46

 List to Textport 46

 Next 46

 Restart 46

 Restore from File 46

 Save to File 46

 Use in Auto Align 46

 overview 2

 Tools and Options 41

Align and Superpose option

 Alignment Options 41

 Atom selection 45

 Atoms to Superpose 43

 Calculate for matched residue only 44

 Choose target molecule 43

 Color by Homology 44

 Cutoff for difference in residue accessibility
 42

 Distance cutoff 42

 Dot Plot Color Ranges 44

 Dot Plot Options 44

 Gap Penalties 44

 List RMS per residue 45

 Maximum accessibility score 42

 Maximum distance score 42

 Move all atoms in molecule 44

 Normalize dot plot scores 44

 Number of residues in window 44

 Output transformation matrix 44

 Plot dendrogram 43

 Plot graph of match scores 43

 Plot RMS per residue 45

 Protein Align Score File 42

 Select matches 43

 Show constraints on dot plot 44

 Superposition Options 43

 Undo All Matches 43

 Update matches when alignment changed 43

 Weight to fix constrained residues in align-
 ment 42

Align and Superpose tool

 Add Gap 42

 Align Sequences... 41

 Align Two Residues... 43

 Alignment Scores... 42

 Alignment Weights 42

 Change Match 43

 Delete Gap 42

 Dot Plot 43

 Finish 45

 Match Options 43

 Match Residues 43

 Options 43

 Plot Dendrogram 45

 Reread MSF 43

 RMS Deviations 44

 Save MSF 43

 Superpose Matched Residues... 43

 Undo All 42

 Undo Last 42

aligning 33

aligning sequences 34, 35

alignment 36

 constraints 39

 manual editing 37

 saving and restoring 37

 scoring 36

alpha carbon torsion convention 76

- Amino Acid Residue tool
 - Finish 23
 - Quit 23
- Amino Acid Selection tool
 - Keyboard 23
 - Non-standard 23
 - Undo Last 23
- amino acids
 - structure definition file 185
- analysis
 - overview 1
- Analyze Domain Structure
 - overview 2
 - Tools and Options 95
- Analyze Domain Structure option
 - One Less Domain 96
 - One More Domain 96
- Analyze Domain Structure tool
 - Create Domain 96
 - Display Cluster 96
 - Finish 97
 - List Domains 96
 - Merge Domains 96
 - Number of Domains 96
 - Options 96
 - Reassign Element 96
 - Reassign Residue Range 96
 - Reread MSF 97
 - Save to MSF 96
 - Undo Domain Edit 96
 - Write Domains to File 96
 - Write Geometry to File 96
- Analyze Secondary Structure 75
 - overview 2
 - tools and options 77
- Analyze Secondary Structure option
 - Alpha Carbon Atoms Only 77
 - Pick Residue 78
 - Pick Residue Range 78
- Analyze Secondary Structure tool
 - Add Hydrogen Bonds 77
 - Assign Secondary Structure 77
 - Calculate Hydrogen Bonds 77
 - Calculate Secondary Structure 77
 - Calculation Options 78

- Color Options 78
- Delete Hydrogen Bonds 77
- Display Hydrogen Bonds 77
- Finish 78
- List to Textport 78
- Read from MSF 78
- Save to MSF 78
- Write to File 78
- analyzing secondary structures 75
- Applications menu
 - Protein Design 139
- Apply Conformation option
 - Position folded fragment 66
- atom-atom repulsion 157
- axial vectors 53, 54
 - average distance 53
 - minimum distance 53
 - same connectivity 54
 - same domain 54
 - scalar angle 54
 - tilt angle/interaxial angle 54

B

- building coordinates 62
- BVALUE 142

C

- C α -C α distance homology 35
 - scoring system 35
- Calculate Accessibility
 - overview 2
 - Tools and Options 82
- Calculate Accessibility option
 - Accessible Area... 82
 - Color Options 82
 - Contact Area 82
- Calculate Accessibility tool
 - Atom Accessibility 82
 - Color by Atom 82
 - Color by Exception 82
 - Color by Fraction Accessible 82
 - Color by Side Chain 82
 - Excluded Zero Accessibility 83

- Finish 83
- List to Textport 82
- Reread MSF 83
- Residue Accessibility 83
- Save to MSF 83
- Write to File 82
- calculating accessibility 79
- calculating contact maps 85
- close contacts 70
- clustering algorithm 94
 - how calculated in QUANTA 94
- commands
 - LIST ATOM 142
 - list location 145
 - modeler model_name 141
 - sys rm -rt directory_name 143
- conservation profiles 29
- contact area 79, 81
 - atom level 81
 - coloring schemes 79
 - definition 81
- contact maps 85
 - calculating 85
 - diagram 86
 - difference contact maps 88
 - difference plot 88
 - distance contact maps 88
 - energy contact maps 89
 - hydrogen bond interactions maps 90
 - interaction-type contact map 89
 - inter-residue C α -C α distances 85
 - inter-residue contacts 87
 - molecule display 88
 - plotting method 86
 - plotting secondary structure elements 87
 - residue type interaction maps 90
 - three major categories 85
 - three types of energy maps 89
- contact plotting method 86
- Create Homology Model
 - overview 2
 - Tools 48
- Create Homology Model tool
 - Change Unknown Structure 48
 - Copy 48

- Copy Matched Residues 48
- Copy Options 48
- Finish 49
- Reread MSF 49
- Save MSF 49
- Select Copy Range 48
- CREBASE program 181, 182

D

- data files
 - conversion 175
 - geometric structure definition file 185
 - PHD format 12, 175
 - protein parameter file 189
 - sequence data format 175
- database 181
 - creation 181
 - creation using the Search Fragment Database
 - utility 179
 - customizing 181
 - search program 109
 - search results 108
 - searching 121
 - structural 109
- database file format 202
- database query 103, 105
- databases
 - overview 1
- deleting residues 22
- dendograms 94
 - definition in footnote 36
- dialog
 - Align and Superpose Options 43
 - Alignment Weights 42
 - Calculate Accessible Area 82
 - Calculate Contact Area 82
 - Change Coloring Ranges 82
 - Change Displayed Map 91
 - Choose an Unknown Structure 48
 - Color Range 44
 - Color Secondary Structure 78
 - Contact Map 90, 92
 - Contact Map Color Ranges 91
 - Cutoff Difference in Average Distance 96

- Database Search Parameters 117
 - Define Coloring Ranges 91
 - Define INTRA-template constraint 114
 - Define phi/psi Dihedrals 134
 - Define Protein Structure Database Query 113
 - Define Template 113
 - Define template name_search 113
 - Delete Template or Constraint 117
 - Display Overlay from Database Search 123
 - Display Selected Fragments 67
 - Enter Number of Domains 96
 - Fold Protein Main Chain 65
 - Fragment Display Mode 120
 - Fragment Modeling Options 67
 - Hydrophobic Profile Options 31
 - Match Option 43
 - Motif Superposition 57, 123
 - Protein Health Options 134
 - Protein Utilities Options 19
 - Score Parameters 42
 - Secondary Structure Analysis Parameters 78
 - Secondary Structure Title 78
 - Select Fragment to Display 118
 - Sequence Database 108
 - Side Chain Torsions 114
 - Specify Protein 104
 - Specify Proteins to Search 116
 - difference contact maps 88
 - directory
 - \$HYD_DMf 180
 - \$MSF_LIB 180
 - \$QNT_ROOT/dmatrix 180
 - \$QNT_ROOT/library 181
 - Display Contact Maps
 - Calculate Contacts... 90
 - Change Displayed Contacts 91
 - overview 2
 - Show Contacts Molecule 1 90
 - Show Contacts Molecule 2 91
 - Tools and Options 90
 - Display Contact Maps option
 - Contact Map Options 92
 - Distance and Energy Difference Maps 92
 - Distance cutoff for energy calculation 92
 - List contacts to file 92
 - Map Colors 91
 - Select one set of residues 91
 - Select two sets of residues 91
 - Show Absolute Values for which molecule 92
 - Show contacts for core residues only 92
 - Show secondary structure on contact map 92
 - Use Distance cutoff in distance difference map 92
 - Display Contact Maps tool
 - Finish 92
 - Only with themselves 91
 - With all of protein 91
 - With non-selected residues 91
 - distance between main chain N and O atoms 160
 - distance contact maps 88
 - distances between sidechain-sidechain and sidechain-mainchain atoms 160
 - dot plot 38
 - attempting dot plots of various window lengths 38
 - path of alignment 38
-
- E**
- Edit Protein
 - Amino Acid Selection 23
 - overview 2
 - Tools and Options 23
 - Edit Protein option
 - Regularization Options 24
 - Edit Protein tool
 - Change Terminal 24
 - Create MSF from Sequence 25
 - Create Sequence 25
 - Delete 24
 - Delete Range 24
 - Disulfide 24
 - Finish 25
 - Insert After 24
 - Insert Before 24
 - Mutate 24
 - Mutate Range 24
 - Regularize 23
 - Renumber Residues 24
 - Use All Hydrogens 24

Use No Hydrogens 24
Use Polar Hydrogens 24
editing proteins 21
energy calculation 75
energy contact maps 89
exclusion range of residues 116

F

FASTA (.aa extension) 193
FASTA search algorithm overview 107
FASTA Sequence Searching 107
feature PDFs 166
file
 \$HYD_EXE/crebase 182
 \$HYD_EXE/msfgen 183
 \$HYD_LIB/database.dat 103, 110, 183
 \$HYD_LIB/motif.geo 122
 \$HYD_LIB/param.par 89
 \$HYD_LIB/pdbsequence.lib 183
 \$HYD_LIB/protein_param.dat 130
 \$HYD_LIB/protein_torsions.dat 65, 81, 130
 \$HYD_LIB/tmplat(noh/pol/all)/*.pdb 73
 \$HYD_LIB/wildcard.da 114
 \$HYD_LIB/wildcard.dat 203
 \$HYD_MSF/dmprep 179
 \$QNT_ROOT/dmatrix/dmfile 180
 \$QNT_ROOT/db/crebase 181
 \$QNT_ROOT/db/crebase/crebase.inp 181
 \$QNT_ROOT/dmatrix/dmfile 66
 database.da 181
 dmfile 180
 dmfile.new 179
 dmlist 179
 dmprep.f 179
 dmsubs.f 179
 HYD_LIB/harvard_torsion.dat 130
 info.log 104
 job_name.ddb 117
 job_name.log 117
 job_name.sel 117
 name_GOR.out 32
 name_momany_predict.out 31
 pdb_master.list 181, 182
 pdbsequence.lib 107, 108, 181

wildcard.dat 117
file formats
 Clustal 11
 EMBL 11
 FASTA 11
 FASTA (.aa extension) 193
 GCG (.gcg extension) 193
 GCG Pairs 11
 GCG Pileup 11
 GCG(Wisconsin) 11
 HAHU 194
 NBRF-PIR (.pir extension) 193
 QUANTA alignment 11
 Swissprot 11
 Swissprot (.sws extension) 193
folding residues 62
fragment searching 63
 Bumps option 64
 useful criteria for choosing a fragment 63

G

gap penalties 36
GAP_PENALTIES_1D 150
GAP_PENALTIES_2D 152
GCG (.gcg extension) 193

H

hbond convention 76
homology
 copying 47
 modeling 47, 61
homology (indicated) 40
homology, structural approaches 149
hydrogen addition 22
hydrogen bonds calculation 75
hydrophobic moments 17
hydrophobicity scales 29

I

inserting residues 22
interaction-type contact maps 89
inter-residue contacts 87

K

Kabsch and Sanders 75
Kabsch and Sanders formula 75
Kraulis, Per 206

L

Lee and Richards accessibility method 80
LIBRARY library_file 124

M

main chain conformation 160
matching residues 39
matching structure motifs 52
Model Backbone
 overview 2
 Tools and Options 64
Model Backbone option
 Accept Fragment 67
 Carry connected residues 66
 Copy fold from another residue range 65
 Display All Fragments 67
 Display Next 67
 Display Previous 67
 Fold in regular structure 65
 Fold to assigned secondary structure type 65
 List Proteins 66
 List Residues 67
 Pick Alpha Carbon 66
 Pick Alpha Carbon Range 66
 Regularize Joins 67
 Reject Fragment 67
 Search Database 67
 Select Display 67
 Spin search side chain conformations 66
 Undo All 66
 Undo Last 66
 User specified phi, psi and omega 65
 with Bumps 67
Model Backbone tool
 Apply Conformation 65
 Build Coordinates 65

Finish 67
Regularization Options 65
Regularize Region 65
Search Fragment Database 66
Undo Last 67

Model Side Chains
 modes of use 71
 overview 2
 Tools and Options 71
Model Side Chains option
 Bump cutoff 73
 Harvard rotamer data file 74
 Hydrogen Bond Cutoff 74
 Protein torsion data file 74
 Radius to display sphere 74
 Spin Increment 74
Model Side Chains tool
 Auto 72
 Build Side Chains 73
 Copy Homologous 73
 Current Residue 72
 Display All 72
 Display Contacts 73
 Display Current 72
 Display Sphere 72
 Finish 74
 Karplus Rotamer 73
 Manually Rotate 73
 Minimize 73
 Next Conformation 74
 Next Residue 72
 Options 73
 Ponders Rotamer 73
 Previous Residue 72
 ReRead MSF 74
 Reset 73
 Save to MSF 74
 Spin Residue 73
 Sutcliffe Rotamer 73
 User Defined 73
model_name.log 143
model_name.top 141, 143
model_name_dir 143
model_name_dir/
 model_name.aln 143

- model_name.B9999nn.pdb 141
- model_name.csr 144
- model_name.D0000nn 144
- model_name.ini 143
- model_name.sch 144
- molecule_name.pdb 143
- molecule_name_dih.Z 144
- molecule_name_ngh.Z 144
- molecule_name_psa.Z 144
- model_name_nn.msf 141
- MODELER**
 - accessing 139
 - command file 144
 - control file 141, 144
 - deleting files 143
 - description 138
 - displaying results 141
 - files 143
 - models
 - adding a disulfide bond 146
 - representing hydrogens 146
 - modifying files to modify models 146
 - optimizing run 140
 - preparing to use 139
 - restraints 147
 - run options 140
 - specifying run parameters 139
 - starting 139
- Modeler**
 - background theory 148
 - structure generation theory 169
- MODELER (introduction) 3**
- \$MODELER_ROOT/exec/ __defs.top. 146**
- modeling 170
 - overview 1
- modeling side chains 22
- modeling the protein backbone 61
- molecule display 88
- MOLECULE msf_name 124**
- MolScript 205**
- MolScript interface 205**
- Motif Database**
 - overview 3
 - Tools and Options 122
- Motif Database option**

- Change Active 122
- Options 123
- Pick Domain 122
- Pick Element 122
- Pick Range 122
- Table Database Results 123
- Motif Database tool**
 - All Overlays 124
 - Auto Database Search 123
 - Clear Display 124
 - Finish 124
 - Next Overlay 124
 - Overlay Database 123
 - Previous Overlay 124
 - Save Structure to Database 124
 - Select Database 124
 - Select Overlay(s) 124
 - Show Overlays 123
- motif log file keyword**
 - LIBRARY library_file 124**
 - MOLECULE msf_name 124**
 - NMATCH number_of_matches 124**
 - TEST molecule_name 124**
- mutating residues 21, 22

N

- NBRF- PIR 194**
- NBRF-PIR (.pir extension) 193**
- NMATCH number_of_matches 124**

O

- optimization level 140
- overview
 - FASTA search algorithm 107

P

- palette**
 - Accessibility 82
 - Align and Superpose 41
 - amino acid selection 23
 - Analyze Secondary Structure 77
 - Contact Maps 90

- Create Homology Model 48
- Domain Analysis 95
- Edit Protein 23
- Fold Motif Database 122
- Fragment Database 66
- Model Backbone 64
- Model Side Chains 71
- Predict Secondary Structure 31
- Profile Analysis 101
- Protein Database 113
- Superpose Folding 54
- PDB
 - retrieving textual information with Protein Information utility 103
 - searching for close sequences 107
- PDB files 142
- PDB format file 143
- PDB master file 181
- pdbsequence.lib file 108
- plotting secondary structure elements 87
- positioning the fold fragment
 - average position 63
 - C- terminus fixed 63
 - N-terminus fixed 63
- PostScript
 - generating 205
- Predict Secondary Structure
 - overview 3
 - Tools 31
- Predict Secondary Structure tool
 - Edit Secondary Structure 32
 - Finish 32
 - GOR Options 32
 - Pick Residue 32
 - Pick Residue Range 32
 - Plot Composition 31
 - Plot Conservation Profile 31
 - Plot GOR Prediction 32
 - Plot Holley/Karplus Prediction 32
 - Plot Hydrophobic Profile 31
 - Plot Momany Prediction 31
 - Profile Options 31
 - Read from MSF 32
 - Save to MSF 32
- probability density functions 154
- basis and feature 165
 - for bond lengths, bond angles and dihedrals 155
 - observed in known protein structures 154
- Profile Analysis
 - overview 3
 - Tools and Options 101
- Profile Analysis tool
 - Align 101
 - Dot Plot 101
 - Options 102
 - Plot Sequence Profile 101
 - Plot Structure Profile 101
 - Read from MSF 102
 - Recalculate Profile 102
 - Save to MSF 102
 - Select Sequence 101
 - Undo Alignment 102
- profiles
 - plotting 100
 - see proteins 99
- program
 - \$HYD_EXE/search 117
 - \$HYD_EXE/search job_name 199
- Protein Design 139
 - overview 1
- Protein Design Alignment palette 33
- Protein Design palette
 - overview 2
 - Run Modeler 139
- protein domains
 - analyzing 93
 - clustering algorithm 94
 - definitions 93
 - dendograms 94
 - how characterized 93
 - importance of the cutoff distance 94
 - loop regions 94
- protein geometry 53
- Protein Health
 - analysis of multiple conformations 128
 - Dunbrack and Karplus rotamer definitions 131
 - exception mnemonics 129
 - presenting the results of the health check 128
 - Tools and Options 130

- uses 127
- Protein Health library
 - Dunbrack and Karplus Rotamer Library 130
- Protein Health option
 - Highlights 135
 - Phi/psi Options 134
- Protein Health tool
 - Buried Hydrophilic and Exposed Hydrophobic Residues 133
 - Buried Polar Atoms 131
 - Chirality 131
 - Close Contacts 133
 - Convert PDB files to CSR 135
 - Display Conformation 135
 - Display Exceptions 134
 - Holes 134
 - List Exceptions to Textport 134
 - List Phi/psi to Textport 135
 - Main Chain Conformation 130
 - Options 134
 - Phi/Psi Plots 134
 - Select Active Residues 134
 - Side chain conformation 130
 - Tabulate MultiConformations 135
 - Undefined Coordinates 130
 - Write Exceptions to File 134
 - Write Phi/psi to File 135
- protein homology modeling 137
- Protein Information
 - overview 3
 - running a query 105
 - Tools 104
- Protein Information tool
 - Maximum crystal resolution 104
 - Output log file name 104
 - Position in database between structure number 104
 - Search for keyword or PDB 104
- protein modeling 61
- protein sequence library
 - pdbsequence.lib file 107
- Protein Utilities
 - Tools and Options 18
- Protein Utilities option
 - Color by Homology 20

- Color By Sequence Properties 20
- Color by Structure Properties 19
- Options 19
- Protein Utilities palette 18, 139
- Protein Utilities tool
 - Atom Information 18
 - Bond Angle 18
 - Center 18
 - Clear ID 18
 - Delete Monitors 19
 - Dihedral 19
 - Distance 18
 - Hydrophobic Moments 19
 - Legend 19
 - Molecule Colors 19
 - Reset View 18
 - Secondary Structure 19
 - Select Active Range 18
 - Set Origin 18
 - Show Monitors 19
 - Smoothed CA Trace 19
 - Torsion Table 19
- proteins
 - editing 21
 - hydrophobic moments 17
 - modeling the backbone 61
 - profile comparison 100
 - profile/sequence comparison 100
 - profiles 99
 - simple representations 17

R

- reference motif 52
- regularization 21, 62
 - when to use regularization 62
- regularizing regions 62
- residues
 - accessibility homology 35
 - accessibility scoring 35
 - environment 35
 - mutating, inserting, and deleting 21
 - rotamers 70
- restraining dihedral angles 163, 164
- restraints 147

- Retrieving Protein Information 103
- RMS Deviation tool 40
- rotamer conformations
 - χ_1 and χ_2 dihedrals 71
- rotamer libraries 70
- Rotamer Library Types 131
- rotamers 70
- Run Modeler 139

S

- scoring 33
- scoring schemes 34
- search command
 - ATOL 198
 - CONS 199
 - CTOL 198
 - DBAS 197
 - DEBUG 199
 - DIST 199
 - DTOL 198
 - INFO 198
 - MKEY 197
 - MNAM 197
 - MRNG 197
 - NHIT 198
 - RESD 198
 - RSLN 198
 - SIDE 199
 - STOL 198
 - TMPL 198
 - WILD 197
- search commands 197
- Search Fragment Database tool
 - Options 67
- searching
 - running a search 199
 - the structure database 110
- secondary structure
 - analyzing 75
 - assignment
 - 3-turn 76
 - 4-turn 76
 - 5-turn 76
 - alpha helix 76
 - beta bulges 77
 - beta strand 76
 - extended conformation 77
 - folded conformation 77
- GOR prediction 29
- Holley/Karplus prediction 28
- homology 35
- homology scoring system 35
- Momany prediction 28
- predicting 27
- representation 54
- segments
 - editing 22
- sequence alignment 54, 149
- Sequence Database
 - FASTA program 107
 - overview 3
 - Tools 108
- Sequence Database tool
 - Change sequence database file 108
 - Enter search sequence 108
 - Read sequence search log file 108
 - Run sequence search 108
- sequence homology 34
- Sequence tool 11
 - Plot Sequence Viewer 14
 - Read Sequence Data File 12, 175
 - Read Sequence/Alignment File 11
 - Remove Sequence 15
 - Write Sequence File 14
- Sequence Viewer 5
 - display 6
 - drawing graphs 7
 - Identifying residues 6
 - Residue IDs and names 6
 - Selecting the viewing area 6
 - Sequence names 6
 - use of mouse 6
- Sequence Viewer icon
 - Activity Selection 8
 - Display Selection 8
 - Expand 8
 - Find It 8
 - Focus 8
 - G Toggle Graph Display 8

- Highlighting 7
- Options 7
- sequences
 - aligning 34
 - changing max number 5
 - data 5
 - data import demo 12
 - data import, export 11
 - saving sequences and alignment between sessions 5
- side chain conformation restraints 162
- side chains
 - automatic modeling 69
 - bump cut-off distance 71
 - centers 201
 - conformation 69
 - modeling 69
 - spinning side chains 71
 - torsion angles 201
- solvent accessibility 79
 - coloring schemes 79
- solvent accessible surface area 79
- solvent sphere 80
 - calculating the solvent surface 80
- spin algorithm 71
- spinning side chains 71
- Structural Database
 - overview 3
 - Tools and Options 113
- Structural Database option
 - Define INTRA-template constraint 114
 - Define template 113
 - Define template name_search 113
 - Restore from MSF 120
 - Save to MSF 120
 - Select sphere 119
 - Select zone 119
 - Side Chain Torsions 114
 - Superpose hits 119
 - Using Distance Constraints 115
 - Using Residue Separation Constraints. 115
 - Write to default MSF name 119
- Structural Database query
 - browsing database search results 113
 - proteins to search 111
 - search constraints 111
 - search parameters 112
 - search templates 111
- Structural Database tool
 - Define constraint between two templates 115
 - Define Query 113
 - Display All Fragments 120
 - Display Next 120
 - Display Option 120
 - Display Previous 120
 - List secondary structure and amino acid wildcards 117
 - List/delete existing template(s) and constraint(s) 117
 - Read Hit Fragments 118
 - Read in database search results 118
 - Read Search Log 118
 - Run Search 117
 - Search Parameters 117
 - Select Display 120
 - Select extra neighboring residues to be displayed 119
 - Select hits to display 118
 - Specify proteins to search 116
 - Superpose Fragments 119
- structure superposition 40
- Superpose Folding Motif
 - overview 3
- superpose folding motifs
 - calculated from axial vectors 54
- Superpose Motif
 - demonstration 59
 - Tools and Options 54
- Superpose Motif option
 - Geometric Criteria of Secondary Structures 57
 - Individual Secondary Structure 57
 - Motif Tables 56
 - Number of matched elements 57
 - Pick Domain 55
 - Pick Element 55
 - Pick Element Range 55
 - Reread MSF 58
 - Save to MSF 58
 - Superposition RMS difference 57
 - Undo Alignment 58

.

Superpose Motif tool

- Align Sequence 58
- All Overlays 58
- Change Activity Tools 55
- Clear Display 58
- Finish 59
- Match Close Residues 59
- Motif Superposition Options 57
- Next Overlay 58
- Overlay Motifs 55
- Previous Overlay 58
- Select Overlay 58
- Superpose Molecule 58

superposing 33

- folding motifs 52

superposing structures 40, 51

Swissprot (.sws extension) 193

SWISSPROT (extension .sws) 195

T

table

- Overlay Motif 56, 123
- Secondary Structure Elements 56, 123
- Torsion Table 19

TEST molecule_name 124

.top file 141

topology library 146

TOPS 144

U

user data

- demo 12

using the motif database 122

utilities 17

V

valid secondary structure types 114

visualizing molecule shape 75

W

wildcard definitions 203

wildcard types 203

Z

z-spacing 80