CMPyMOL: A Tool for Protein Contact-Map Analysis

Venkatramanan Krishnamani*

Address: 5-632 Bowen Science Building

Dept. of Molecular Physiology and Biophysics,

University of Iowa, Iowa 52242, US

Telephone: (949) 202 9466

E-mail: venky.krishna@icloud.com

Keywords: Contact Map, PyMOL, Software

Abstract

Contact-maps are reduced 2D representation of the 3D spatial configuration of a protein. Many valuable structural features like secondary structures, inter- and intra-protein interactions, interacting domains, etc., can be readily identified from these maps. However, it is not straightforward and intuitive to reckon the spatial organization of the contact regions from reduced representation. The CMPyMOL software attempts to bridge this gap as an interactive graphical tool for protein contact-maps that interfaces with PyMOL for 3D visualization. Importantly, CMPyMOL helps understand the functional importance of contacts by providing visual overlays of various structural and biochemical properties of a protein on top of its contact-map.

Availability:

CMPyMOL is freely available for Linux and Mac OS X platforms. The source code and user guide can be downloaded from http://emptyewer.github.io/CMPyMOL. The source code is licensed under The MIT License.

1 Introduction

A contact-map of a protein is a 2D matrix of pairwise inter-residue distances, typically calculated over distances between Cα atoms subject to an arbitrary maximum threshold. By construction this distance matrix is square and symmetrical. Contact-maps have been traditionally used to compare two protein structures/conformations (Caprara et al., 2004), protein-protein-interactions (de Melo et al., 2006, Melo et al., 2007), protein folding (Park

et al., 2000, Vendruscolo and Domany, 2000), structure prediction (Di Lena et al., 2012) and even reconstruction of the protein's 3D structure (Duarte et al., 2010).

Contact-maps capture high resolution, residue-level information regarding local conformations such as α-helices and β-sheets, and non-local interactions like inter-domain interactions. In this respect, contact-maps are a lossless representation of structural information (except for chirality). However, essential biochemical information such as the residue type and the properties associated with it are lost during a contact-maps' construction. The interaction type of a particular contact point, such as hydrophobic interactions, salt bridges and hydrogen bonds, etc., can be crucial in understanding protein structure and function. A manual assignment to keep track of the residue number, the residue-type from the protein sequence and its spatial location can quickly become cumbersome and unmanageable.

CMPyMOL supplements a contact-map analysis by interfacing with the powerful 3D visualization capabilities of PyMOL (DeLano, 2002) and provides visual overlays of structural and biochemical properties of the amino acid residues. Secondary structure (using STRIDE (Frishman and Argos, 1995, Heinig and Frishman, 2004)), charge-charge interactions, hydrophobic interactions, B-factor and custom selected residue pair interaction sites (see **Figure 1**) are the currently available overlays in CMPyMOL. The program also calculates pairwise residue interaction heat-map and residue-wise contact density plots as an alternate representation of the contact-map data.

CMPyMOL provides a much needed add-on to the PyMOL software package, a tool which is typically a built-in part of other molecular visualization programs, such as VMD (Humphrey et al., 1996). There exists at least one other interactive tool for visualizing contact-maps, but it is limited in terms of displaying biochemical and structural information

on the contact-maps and does not support multi-frame PDB trajectories (Vehlow et al., 2011). CMPyMOL is intended to replace an existing PyMOL plugin, Contact Map Visualizer (available from http://www.pymolwiki.org), that was co-developed by the author (VK) in collaboration with Thomas Holder (Schrödinger, LLC).

2 Software Functionality

Launching CMPyMOL invokes the PyMOL executable and generates a contact-map (for a specified cut-off distance) for a given PDB file. Visualizing multi-frame PDB trajectories are also supported from Version 1.1 onwards. The user is provided with an option to calculate the distance map between Cα or Cβ atoms of residue pairs and the desired cutoff distance. The main program window (**Figure 1**) displays the contact-map in gray scale, shaded according to the distance between the pair of Cα atoms. CMPyMOL is developed using the Python programming language and the open source libraries Numeric python, Python imaging library and Matplotlib.

2.1 Substructure Selection

Contact points can be selected individually (left-click) or as contact regions (rectangular selection by left-click and drag) interactively over the contact-map. When a substructure of interest is selected, the corresponding residues are synchronously highlighted in the PyMOL window. It should be noted that CMPyMOL only highlights the residues that are in contact and within the current selection. By construction, the lower and upper half of the contact-map separated by the main diagonal are symmetric, hence selecting contacts on either side of the diagonal draws a corresponding selection rectangle on the other side.

2.2 Overlays and Plots

A novel feature in CMPyMOL are the overlays of structural and biochemical properties that highlights the nature of interaction of each contact point. This allows clear distinction, discovery and classification of contacts of a specific biochemical/structural type. For example, turning on "Charged Interactions" displays a *blue* overlay on all the contacts points where two charged residues are in proximity, i.e. in contact (while "Hydrophobic Interactions" as displayed as an *yellow* overlay). "B-factor" enables a *cyan* highlight of residues that have a Debye-Waller factor (Debye, 1913, Waller, 1923) above a certain cutoff (specified by the slider, **Figure 1**) and that are in contact. Additionally, users can select any pair of amino-acids, listed on the right, to display an *orange* overlay on contact points where the selected amino-acids are in contact with one another. This is a powerful tool for quickly and efficiently identifying specific interactions types and simultaneously visualizing their spatial orientation.

The "Pairwise Heat-map" and "Contacts Histogram" (**Figure 1 and Figure 2**, buttons on the left of the contact-map), calculates and plots the number of pairwise residue contacts and contact density of each residue, respectively. Of these maps, the contact density map is interactive, mouse selections of a particular contact density highlights the location in the PyMOL window. When a multi-frame PDB trajectory is loaded into the software, the user can choose to view either the contact-map or the cumulative variance contact-map calculated for all the frames starting from frame 1 to the current frame selection.

2.3 Limitations

Currently, CMPyMOL only supports importing locally available PDB file-formatted files. Since each pixel of the contact-map image represents an interacting residue pair, the number of residues of a protein that can be comfortably displayed on a computer screen is limited to approximately 1000-1500 (varies by screen resolution). The software is currently not supported on Windows operating system.

3 Conclusion

CMPyMOL integrates 2D contact-maps augmented with biochemical information and powerful 3D Visualization of PyMOL. This provides an intuitive platform for simultaneously exploring protein interaction sites and its 3D structure.

4 Future Directions

Next iteration of the software will also add support for importing additional file formats, importing from on-line PDB database and the ability to export the history of substructure selections for post-processing.

Acknowledgment

I would like to thank my postdoctoral advisor Dr. Markus Deserno at Carnegie Mellon University for his support and insightful suggestions.

Funding: Volkwagen Foundation for supporting this project within the framework of the

program "New Conceptual Approaches to Modeling and Simulation of Complex Systems"

Disclosure Statement

No competing financial interests exist.

References

Caprara, A., Carr, R., Istrail, S., Lancia, G., and Walenz, B. (2004). 1001 Optimal PDB Structure Alignments: Integer Programming Methods for Finding the Maximum Contact Map Overlap. *J Comput Biol*, 11(1):27–52.

de Melo, R. C., Lopes, C. E. R., Fernandes, Jr, F. A., da Silveira, C. H., Santoro, M. M., Carceroni, R. L., Meira, Jr, W., and Araújo, A. d. A. (2006). A Contact Map Matching Approach to Protein Structure Similarity Analysis. *Genet Mol Res*, 5(2):284–308.

Debye, P. (1913). Interferenz von Röntgenstrahlen und Wärmebewegung. *Annalen der Physik*, 348(1):49–92.

DeLano, W. (2002). The PyMOL Molecular Graphics System.

Di Lena, P., Nagata, K., and Baldi, P. (2012). Deep Architectures for Protein Contact Map Prediction. *Bioinformatics*, 28(19):2449–57.

Duarte, J. M., Sathyapriya, R., Stehr, H., Filippis, I., and Lappe, M. (2010). Optimal Contact Definition for Reconstruction of Contact Maps. *BMC Bioinformatics*, 11:283.

Frishman, D. and Argos, P. (1995). Knowledge-based Protein Secondary Structure Assignment. *Proteins*, 23(4):566–79.

Heinig, M. and Frishman, D. (2004). STRIDE: A Web Server for Secondary Structure Assignment from Known Atomic Coordinates of Proteins. *Nucleic Acids Res*, 32(Web Server issue):W500–2.

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: Visual Molecular Dynamics. *J Mol Graph*, 14(1):33–8, 27–8.

Melo, R. C., Ribeiro, C., Murray, C. S., Veloso, C. J. M., da Silveira, C. H., Neshich, G., Meira, Jr, W., Carceroni, R. L., and Santoro, M. M. (2007). Finding Protein-Protein Interaction Patterns by Contact Map Matching. *Genet Mol Res*, 6(4):946–63.

Park, K., Vendruscolo, M., and Domany, E. (2000). Toward an Energy Function for the Contact Map Representation of Proteins. *Proteins*, 40(2):237–48.

Vehlow, C., Stehr, H., Winkelmann, M., Duarte, J. M., Petzold, L., Dinse, J., and Lappe, M. (2011). CMView: Interactive Contact Map Visualization and Analysis. *Bioinformatics*, 27(11):1573–4.

Vendruscolo, M. and Domany, E. (2000). Protein Folding Using Contact Maps. *Vitam Horm*, 58:171–212.

Waller, I. (1923). Zur Frage der Einwirkung der Wärmebewegung auf die Interferenz von Röntgenstrahlen. *Zeitschrift für Physik A Hadrons and Nuclei*, 17(1):398–408.

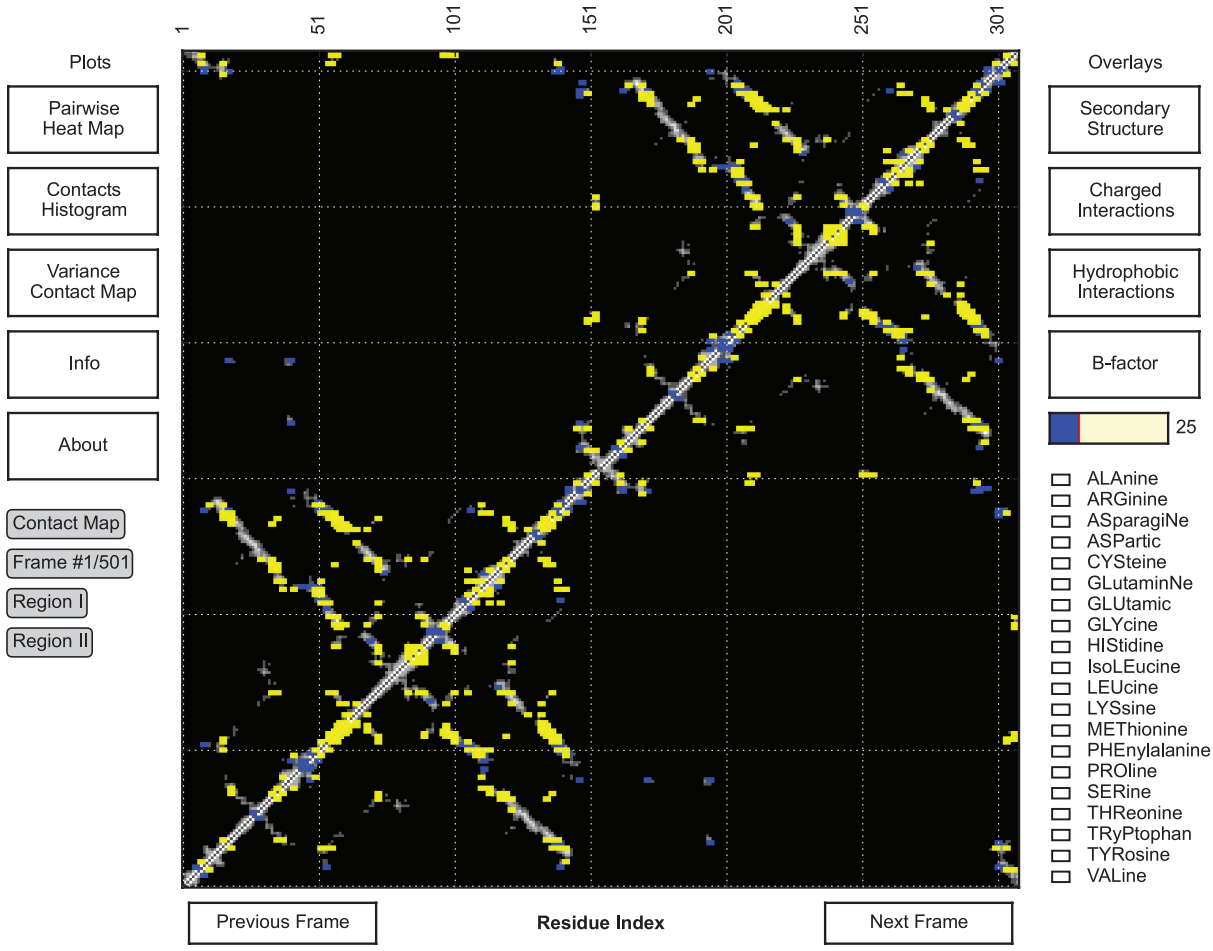


Figure 1: Screenshot of CMPyMOL Main Window: contact-map (center) of a dimer of CCMV capsid protein (PDB code: 1ZA7), calculated as the distance between all pairs of Cα atoms with a cutoff threshold of 10Å. The locations on the contact map where charge-charge interactions (blue dots) and hydrophobic interactions (yellow dots) in the given structure is overlaid. Here, an interaction is defined by the cut-off described above. Also visible are various buttons for displaying other structural/biochemical information as an overlays (on the right) and statistical plots (on the left). CMPyMOL also allows for calculating contact map trajectories and their cumulative variance of the contacts when a multi-frame PDB file is provided (not shown).

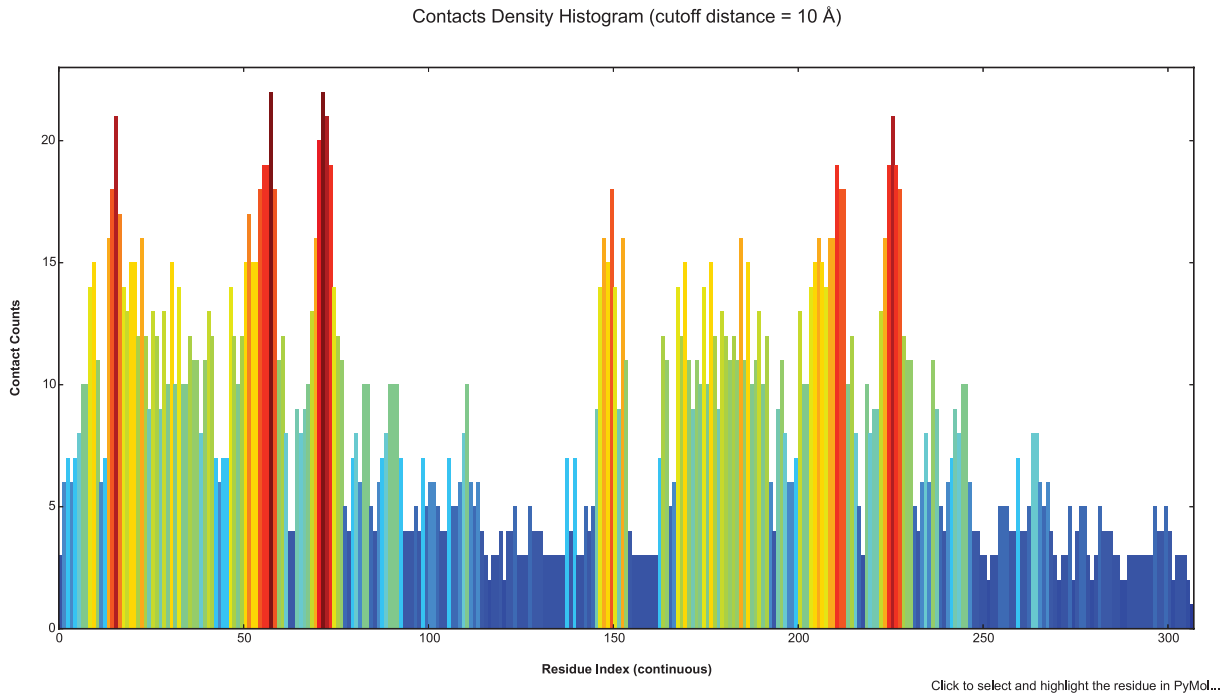
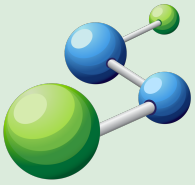


Figure 2: Contact density histogram for CCMV capsid protein (PDB code: 1ZA7). The plot shows the packing density of each residue of this protein and the bars are colored by a color map that scales from the lowest (blue) to highest (red) contact density. This histogram is interactive and clicking on a particular bar of the histogram updates the PyMOL viewer to highlight the corresponding residue and its neighboring residues within a user-defined cutoff.



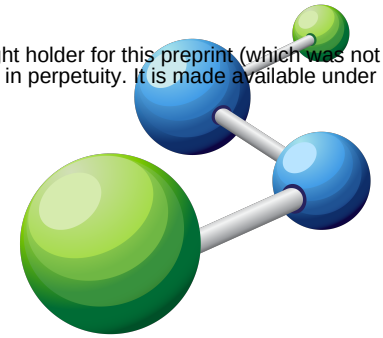
CMPyMOL 2.0

Contact Map plugin for Python Molecular Viewer

Dr. Venkatramanan Krishnamani

HTTPS://GITHUB.COM/EMPTYEWER/CMPYMOL/RELEASES

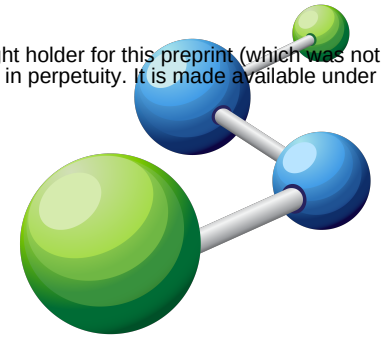
First printing, January 2016



Contents

III Part Two

Part One



1. Installation Instructions

1.1 Pre-compiled Binaries

1.1.1 Download Link

Platform-specific compiled binaries (*Mac OS X and Windows*) of **CMPyMOL** can be downloaded from the below URL.

```
https://github.com/emptyewer/CMPyMOL/releases
```

1.1.2 Prerequisites for pre-compiled binaries

```
# 1. Download and Install PyMOL (according to your platform)
https://www.pymol.org/
```

1.1.3 Mac OS X Compatibility

```
Mac OS X (10.10+) Yosemite and above
```

1.1.4 Windows Compatibility

```
64-bit or 32 bit Windows 7 and above.
Note that CMPyMOL itself is a 32-bit software.
```

1.1.5 Linux Compatibility

```
64-bit Linux Binaries not available for this release.  The next release will
include the linux binaries.
```

R Other machine specific binaries will be provided upon request.

1.2 Running from Source

CMPyMOL source can be downloaded fromt he following URL.

```
https://github.com/emptyewer/CMPyMOL
```

The following libraries are neceassary to run CMPyMOL from source.

R All necessary libraries and software dependencies (except PyMOL) are included in the compiled binaries, this requirement is only for running CMPyMOL from source

1.2.1 Pre-requisite Softwares

1. Python 2.7
2. PyMOL in system $PATH (for all platforms)
3. STRIDE secondary structure prediction software properly configured to be in $PATH

```
http://webclu.bio.wzw.tum.de/stride/
```

R Stride for Mac OSX and Windows is included in the precompiled binary.

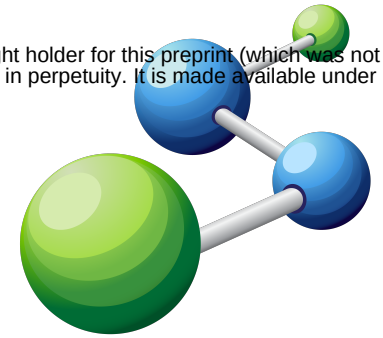
1.2.2 Pre-requisite Python Libraries

1. PyQT4
2. matplotlib
3. xmlrpclib
4. numpy

1.3 Open Source License

```
The MIT License (MIT)

Copyright (c) 2016 Venkatramanan Krishnamani

Permission is hereby granted, free of charge, to any person obtaining a copy
of this software and associated documentation files (the "Software"), to deal
in the Software without restriction, including without limitation the rights
to use, copy, modify, merge, publish, distribute, sublicense, and/or sell
copies of the Software, and to permit persons to whom the Software is
furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all
copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR
IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY,
FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE
AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER
LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM,
OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE
SOFTWARE.
```

Part Two



2. Basic Usage

Currently both single frame and multi–frame PDB files are supported for analysis by CMPyMOL.

2.1 Launch CMPyMOL

CMPyMOL can be launched after installation by clicking on the executable with an icon  (appropriately for Windows and Mac OS X platform). To launch from source simply type `python CMPyMOL_2.0.py` in machine appropriate command line shell.

If PyMOL executable is not automatically located by CMPyMOL, a pop-up will request the location of PyMOl executable. Select either the location of `MacPyMOL.app` (for Mac OS X) or `pymol.exe` (for Windows). The user has to select the location of PyMOL executable only once, after which the path is remembered. In subsequent launching of CMPyMOL if PyMOL is not automatically launched, check the troubleshooting section **??**.

2.2 CMPyMOL Interface

After lanching CMPyMOL, the user will be presented with a window similar to Figure 2.1. This interface provides all the functionalty of the software with no additional pop-ups that distract the user.

The overall organization of the interactive elements in this interface is such that all overlay's that are displayed on top of contact map are on the right side. The parameters for calculating the contact map and other information are provided on the left.

All the maps (contact map, heat map of pairwise amino acid contacts and contact density histogram) are plotted as tabs in the central widget. These tabs will be added to the CMPyMOL interface when a PDB file is loaded. The functionatily of each interactive element will be discussed further in the following sections.

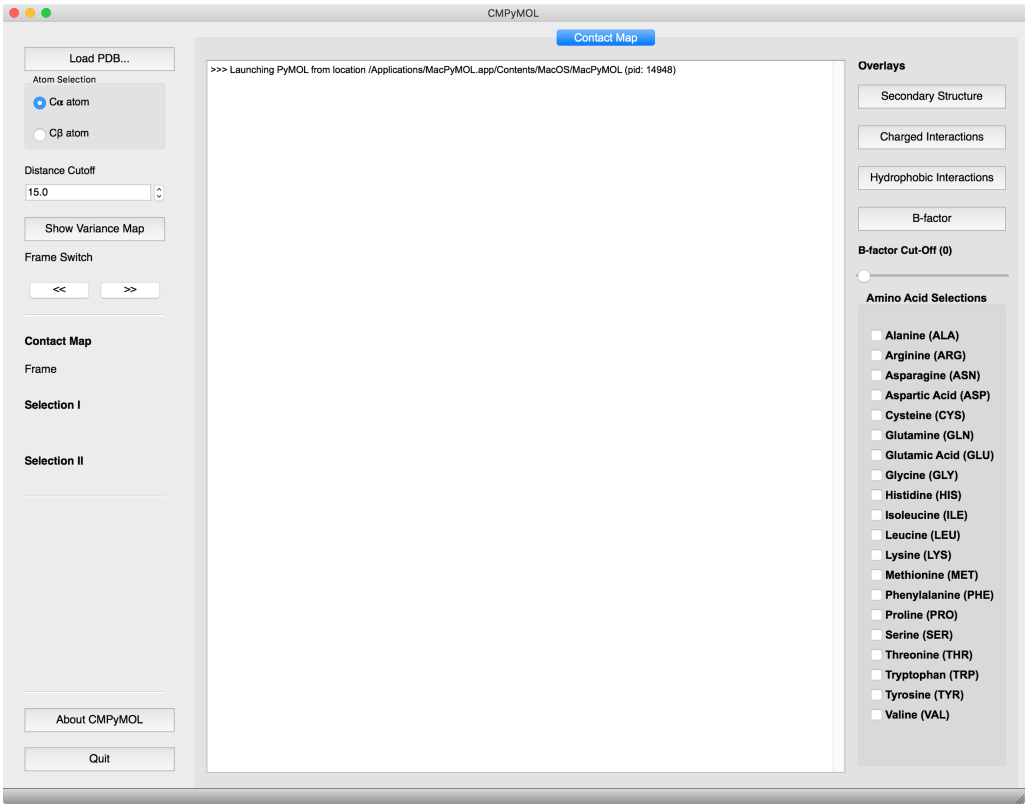


Figure 2.1: Main Interface of CMPyMOL.

2.3 Parameters to Calculate Contact Map

A protein contact map represents the pairwise distance between all possible amino acid residue pairs of a three-dimensional protein structure. These distances are typically calculated between either `Cα or Cβ` atom. The `cut-off` distance eliminates all the pairwise distances that are larger than that value.

The user has the choice to calculate the contact map based on specific parameters (Figure 2.2). The choice of `Cα or Cβ` atom as the basis atom and the cut-off distance (defaults to 15) for calculating the contact map will be can be customized by the user.

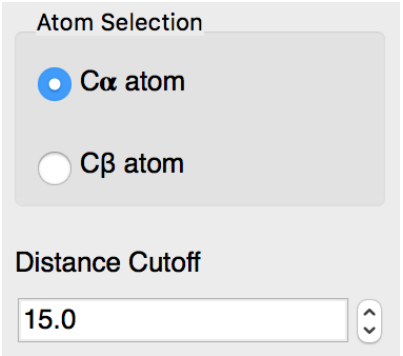


Figure 2.2: Parameters for Contact Map Calculation

R The parameters to calculate contact maps can be set by the user before or after loading the PDB for on-the-fly updates of maps and overlays.

2.3.1 Load PDB file

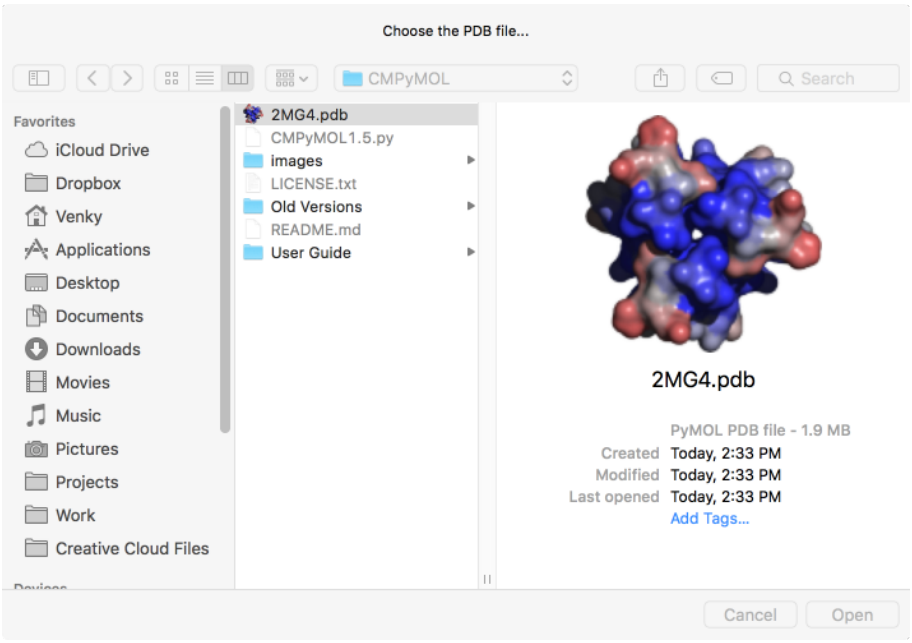


Figure 2.3: Pop-up dialog for file selection.

Either a single frame PDB file or a multi-frame PDB from MD simulation trajectory or NMR trajectory can be loaded at this step. The correct format of the PDB file is provided in Section 2.3.2 for reference. Other structure formats are currently not supported.

Click on the Load PDB... to invoke a pop-up to locate the PDB file. A file dialog (Figure 2.3) will facilitate this choice. Throughout this guide PDB ID: 2MG4 will be used for illustration. 2MG4 is an NMR structure of an artifically designed protein that is a symmetric protein homodimer [1].

After loading the PDB file, there will be two windows presented to the user as shown in Figure 2.4 and 2.5.

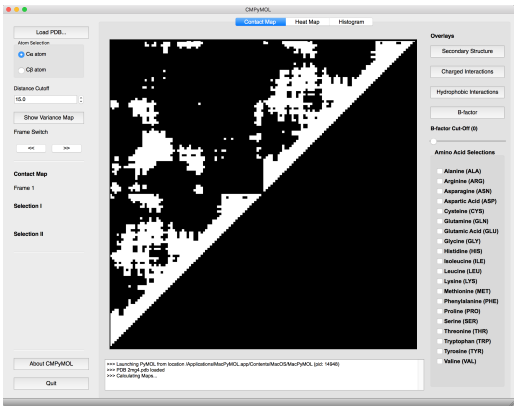


Figure 2.4: CMPyMOL window.

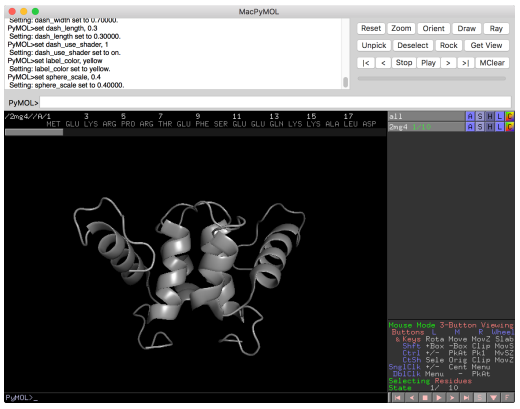
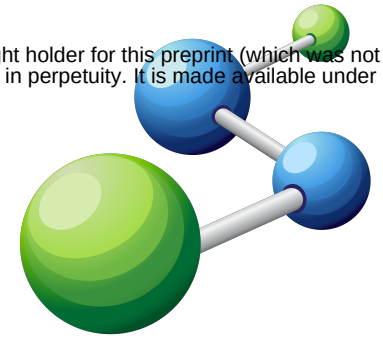


Figure 2.5: PyMOL window

2.3.2 PDB Format

```
MODEL X
.
.
.
ATOM ...
ATOM ...
ATOM ...
.
.
.
ENDMDL
```

Part Two



3. Functionality

The overlays in CMPyMOL enriches the interpretation of contact maps.

3.1 Contact Map

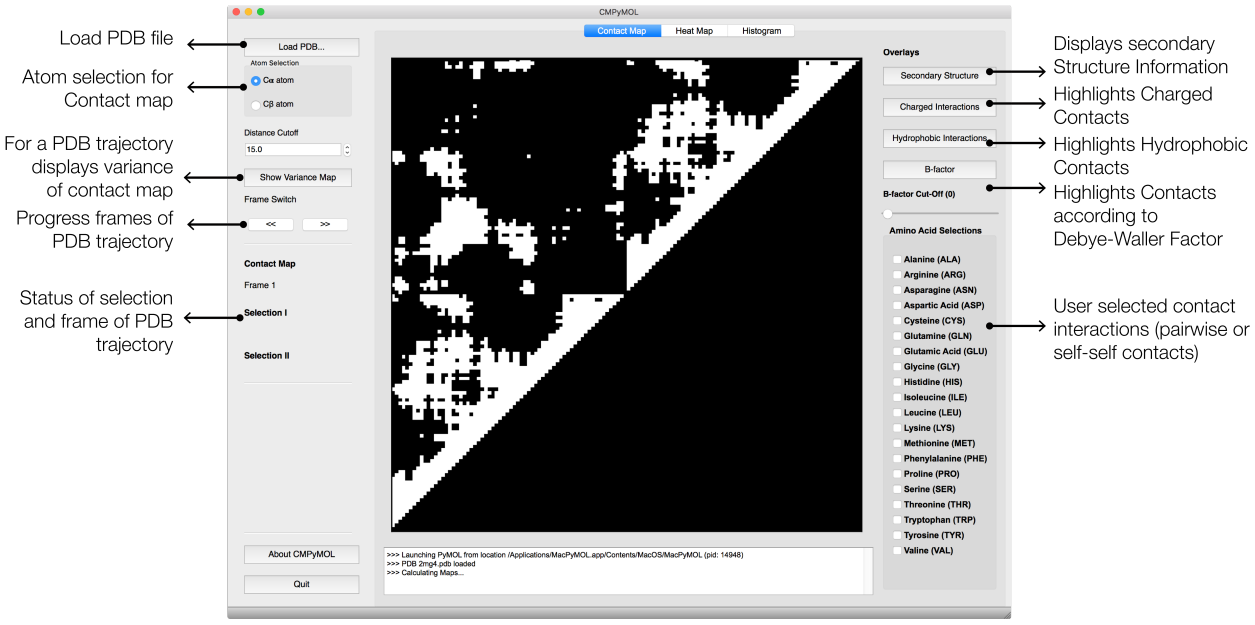


Figure 3.1: Main CMPyMOL Window and Its Functionality

The main window of CMPyMOL (Figure 3.1) provides controls for all the selection, overlay and plots to analyze contact maps. The overlays (the toggle buttons on the right of the contact map) superpose chemical and structural information on top of the contact map when activated. The plots (buttons on the left side of the contact map) pops open a new window that provides an overview of

the nature of contacts.

3.1.1 Variance Contact Map

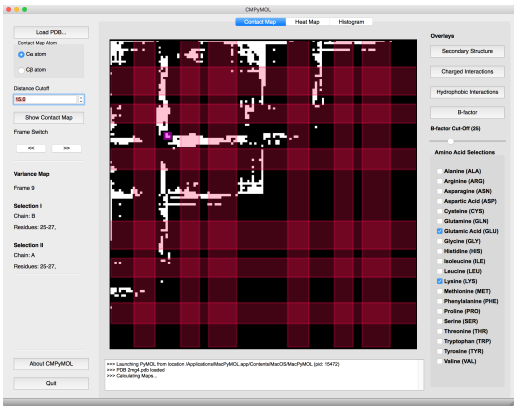


Figure 3.2: Variance Contact Map. Showing the regions of highest flexibility.

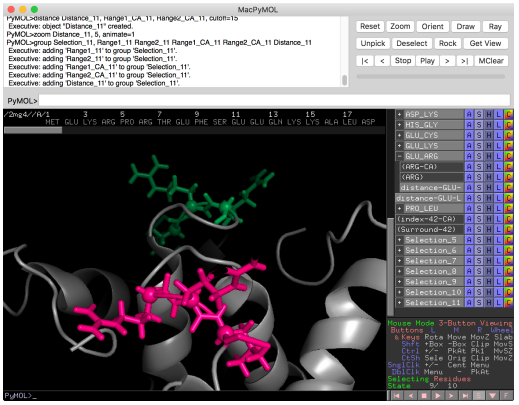


Figure 3.3: Selecting a high variance region in the intra-subunit quadrant reveals in the PyMOL window that the region is a flexible loop.

Variance map can be activated by clicking on the button 

. The red rectangles in figure 3.2 overlays the secondary structure information of the PDB (more in section 3.3.1).

The PDB loaded as an example here is a NMR structure with multiple-frames (10 models). When a multi-frame PDB is loaded, the "Variance Contact Map" button is enabled and it allows for calculating the variance in contact distance for a residue pair. Using 

buttons the user can progress through the different models. Such frame switch will update all the maps in CMPyMOL along with updating the model itself in the PyMOL window.

This allows for readily identifying flexible regions of the protein. In figure 3.2 a selection highlights the intra-protein flexible region with high variance. In the PyMOL window it can be verified that indeed the region of high variance is on a loop (Figure 3.3).

3.2 Selections

Selecting a particular region by clicking-n-dragging on a portion of the CMPyMOL generated contact map (shown as a magenta selection box Figure 3.4) highlights the corresponding region in the PyMOL window (Figure 3.5). Since contact map is a pairwise interaction matrix, the two interacting regions are colored hotpink and limegreen. The atom based on which the distances for the contact map are highlighted as "spheres".

3.3 Overlays

Overlays in CMPyMOL provide a intuitive superposing of chemical and structural information on top of the contact map. This allows of correlating interacting residues with its type (chemical or structural).

3.3.1 Secondary Structure

The button 

toggles the overlaying the secondary structure of protein onto the contact map (Figure 3.6).

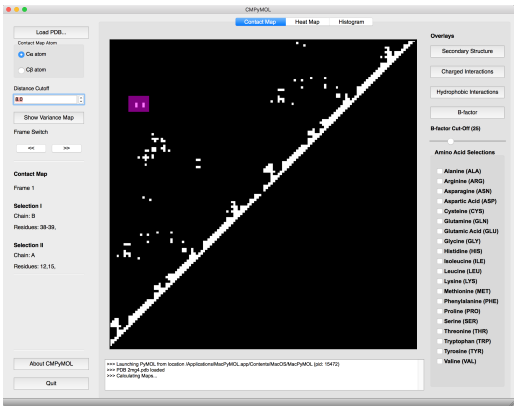


Figure 3.4: Selection in CMPyMOL window.

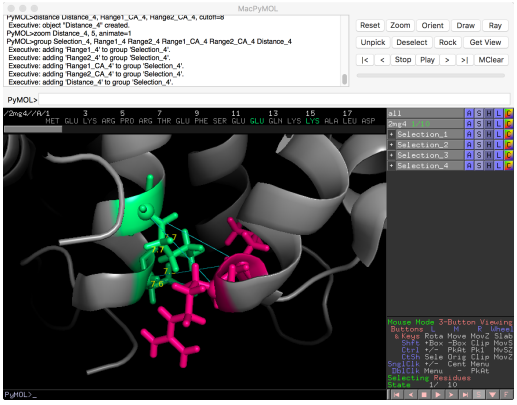


Figure 3.5: Focus on the selected region in PyMOL window.

The secondary structure overlay superposes α-helical and β-sheet as red and green translucent rectangles, respectively. Notedly, the selection from the last section (Section 3.2) is in a α-helix – α-helix interaction, which can be confirmed by the selection highlighted in PyMOL window (Figure 3.5).

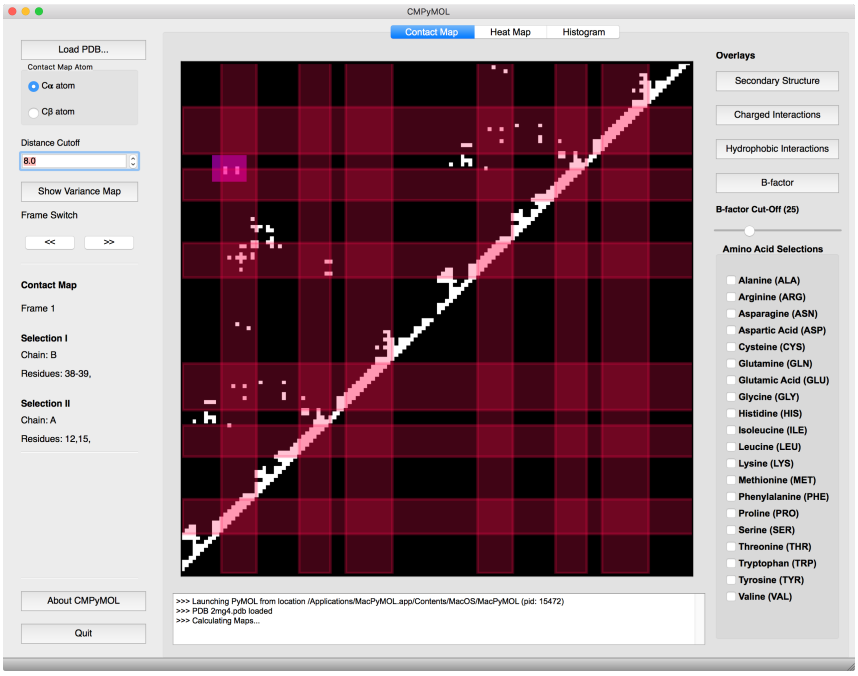
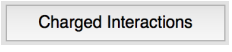


Figure 3.6: Secondary structure overlay on top of contact map.

3.3.2 Charged Interactions

Similar to secondary structure overlay, by clicking on the charged interaction toggle button 

draws a overlay that highlights contact points where two charged residues are located within the cut-off distance specified in Section 2.3 (Figure 3.7). These charged interaction points are highlighted as marine pixels on top of the contact map.

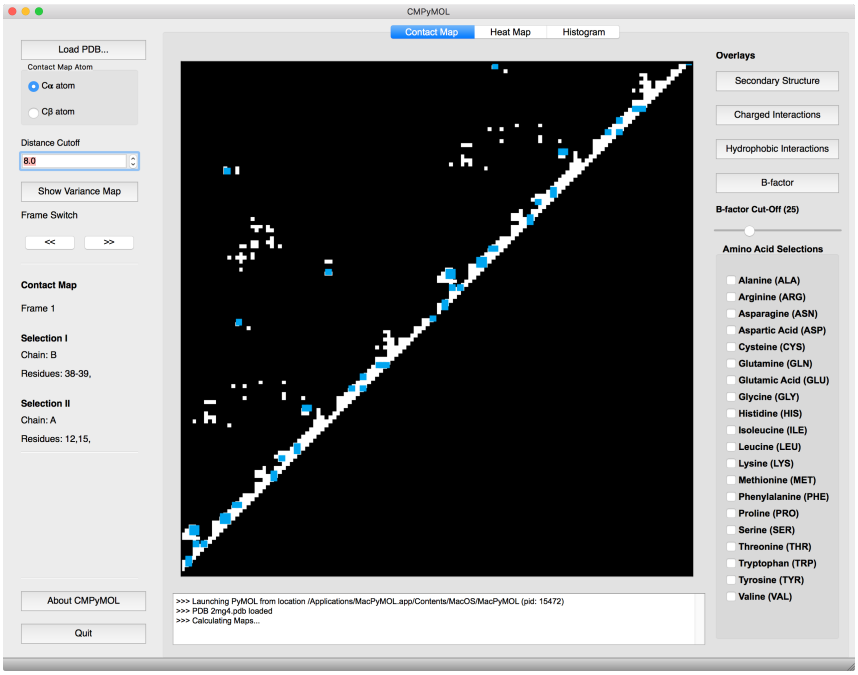


Figure 3.7: Charge-charge interactions highlighted on top of contact map.

3.3.3 Hydrophobic Interactions

Hydrophobic interactions overlay can be toggled by clicking the button Hydrophobic Interactions. This highlights the points of contact map where two hydrophobic residues are coming in contact. These interaction points are highlighted as yellow over the contact map (Figure 3.8).

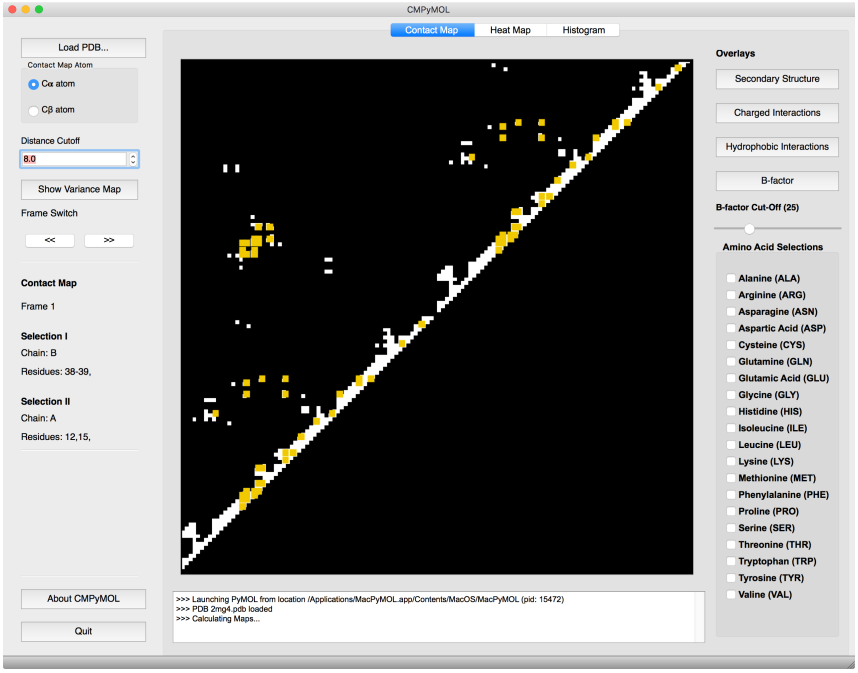


Figure 3.8: Hydrophobic interactions highlighted on top of contact map.

3.3.4 B-factor Overlay

In a similar fashion B-factor (if those values are listed in the PDB) highlights the points with a larger value than set with the slider below the B-factor button (in the case shown in Figure 3.9, the highlighted contact points have a b-factor larger than 25). These interaction points are highlighted as cyan over the contact map.

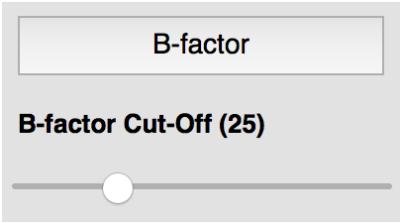


Figure 3.9: Parameters to draw an overlay to display B-factor. The contact points with b-factor larger than cutoff (in brackets) is colored in cyan.

3.3.5 User–defined Pairwise Aminoacid Interactions

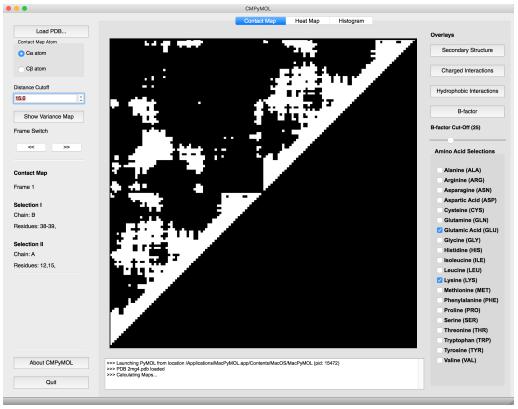


Figure 3.10: User selected interaction between Glutamic Acid and Lysine residues with a a cutoff distance of 15.

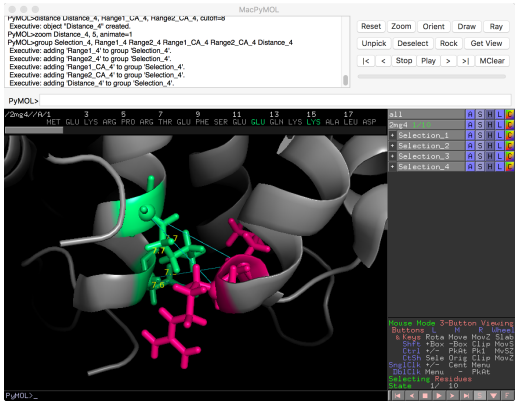


Figure 3.11: The residues corresponding with the selection (Figure 3.10) are highlighted in the PyMOL window.

In addition to the above overlays, the user can choose to highlight interaction points of two specific aminoacids. In Figure 3.10, the interactions of Lysine and Glutamic acid residues are highlited in the PyMOL window (Figure 3.11).

3.4 Plots

Plots in CMPyMOL provide statistical analysis of the contact points.

3.4.1 Pairwise Heatmap

This heatmap counts the number of pairwise contacts of a given aminoacid to the rest of the other aminoacids (Figure 3.12). The order of aminoacids in this plot are arranged according to their hydrophobicity. The color scale shows the number of each pairwise contacts in the protein. The plot is interactive and clicking on any “box” will correspondingly update the pymol window corresponding to the selection.

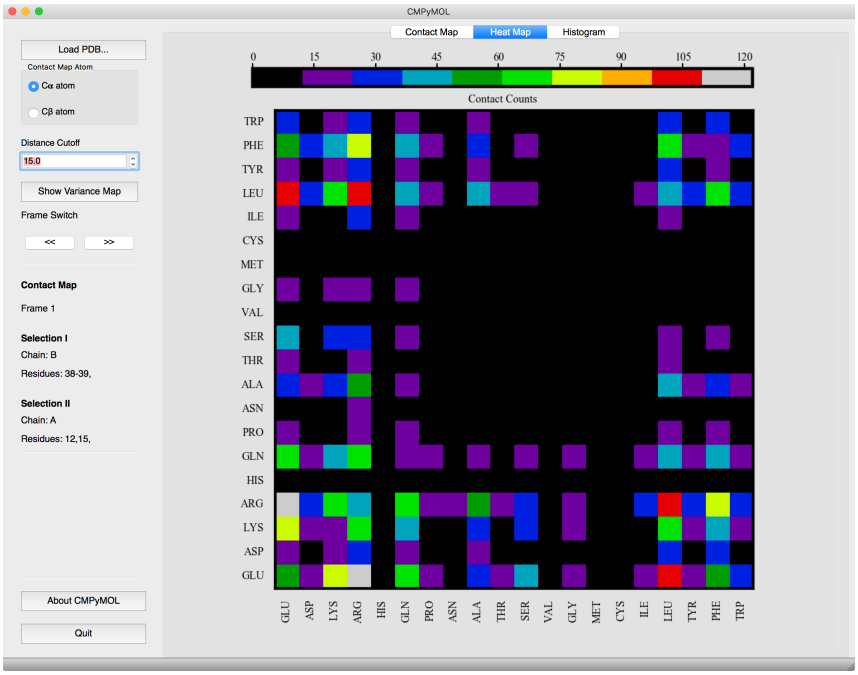


Figure 3.12: Heatmap of pairwise residue-residue interaction map. The residues are listed according to increasing hydrophobicity

3.4.2 Contact Density Histogram

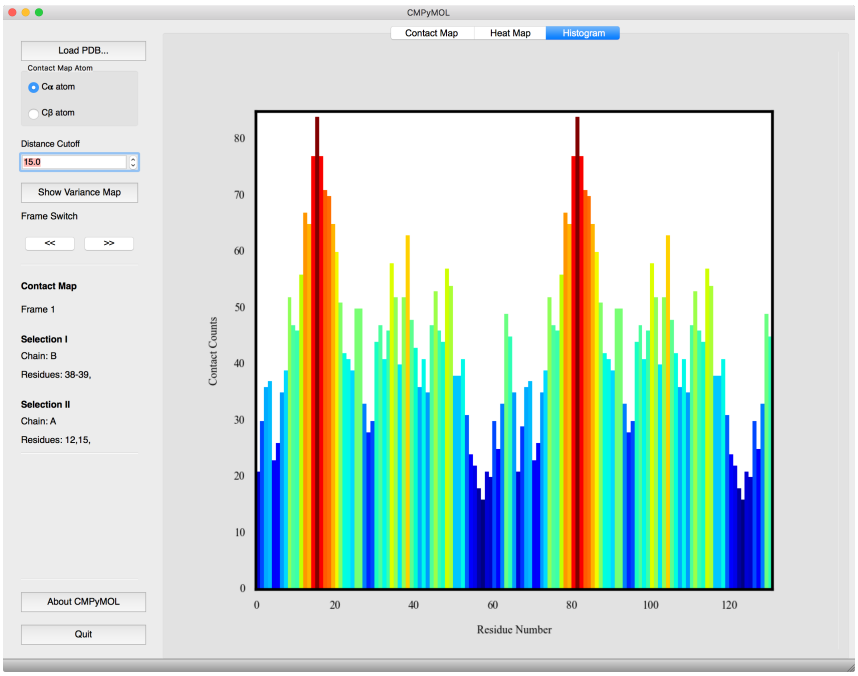
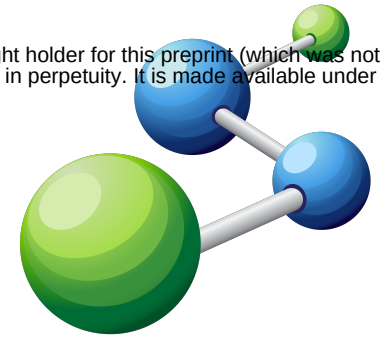


Figure 3.13: The residue-wise density of contacts in the PDB.

The contact histogram plot, graphs the density of contacts with respect to residue position (Figure 3.13). This plot is interactive and clicking on a particular bar selects the corresponding residues and the surrounding contacts in the PyMOL window.

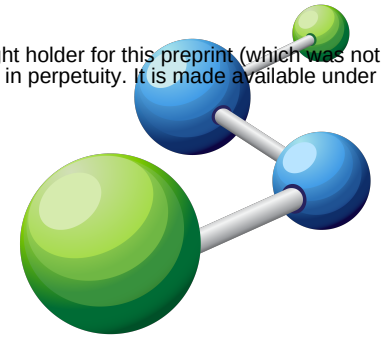


Bibliography

Books

Articles

[Mou+15] Yun Mou et al. “Computational design and experimental verification of a symmetric protein homodimer”. In: *Proceedings of the National Academy of Sciences* 112.34 (2015), pages 10714–10719 (cited on page 13).



Index