PROTMAP2D v. 1.2.2

User’s Manual

PROTMAP2D is an application for the analysis of protein structures through two-dimensional contact maps and distance maps.

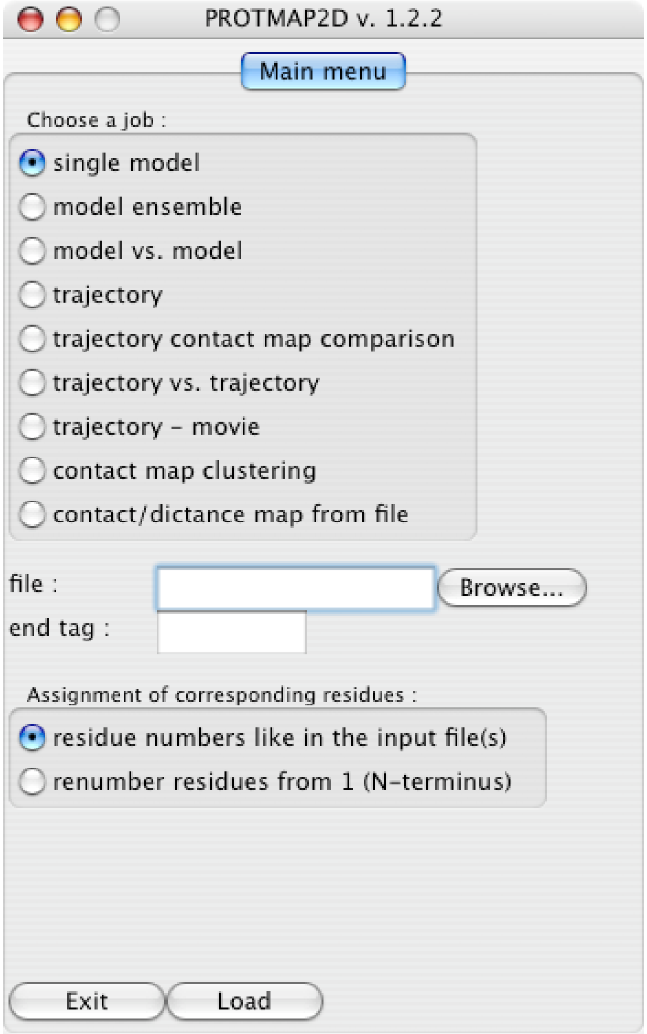
To obtain a contact map with **PROTMAP2D**, the user has to follow four main steps:

1. Select the type of analysis to be performed (hereafter referred to as ‘job’) and define the residue position extraction way

2. Load the input file(s) and have them pre-processed by the program

3. Provide job-specific parameters

4. Analyze the graphical visualization of the results, save the map and statistics to an output file.



Choose a job :

● single model
○ model ensemble
○ model vs. model
○ trajectory
○ trajectory contact map comparison
○ trajectory vs. trajectory
○ trajectory - movie
○ contact map clustering
○ contact/dictance map from file

file :

end tag :

Assignment of corresponding residues :

● residue numbers like in the input file(s)
○ renumber residues from 1 (N-terminus)

1. SELECT THE TYPE OF A JOB

1.1 Single model (requires one PDB file)

Calculate and visualize a contact map or a distance map for one protein structure.

1.2 Model ensemble (requires one PDB file with multiple models)

Analyze the frequency of different contacts in a set of models of identical sequence and length, e.g. an NMR ensemble or a cluster of decoys obtained from *de novo* folding analysis.

1.3 Model vs. model (requires one PDB ensemble or two PDB files)

Compare contacts in two different models of the same protein structures, e.g. a theoretical prediction vs. the structure experimentally determined or models obtained under different conditions, two close homologues etc. (requires one PDB ensemble or two PDB files). The models don't have to be of the same length, but the corresponding residues must have the same number in all models

1.4 Trajectory (requires one PDB file with a series of models separated by same tag)

Analyze the similarity of contacts for models from an ordered series, e.g. snapshots from a simulation by Molecular Dynamics. The first model (presumably a native structure) will be numbered by "0" and regarded as a reference, to which all the other models will be compared. The authors of PROTMAP2D use this option for the analysis of unfolding simulations, which start from the native structure. It facilitates clustering of models based on the similarity of their contact maps.

1.5 Trajectory contact map comparison (requires one or two PDB files, each with a series of models)

Calculate the pairwise contact similarity for models from two trajectories (all against all) and plots it as a function of the model number. It facilitates the identification of similar structures that appear in independent trajectories, e.g. recurring folding intermediates.

1.6 Trajectory vs. trajectory (requires two PDB files, each with a series of models)

Calculate the pairwise contact similarity for equivalent models from two trajectories of the same size (t1-mod.1 vs t2-mod.1, t1-mod2 vs t2-mod.2, etc) and plots it as a function of a residue number. The result shows to which extent each contact was observed in the equivalent frames in the two trajectories.

1.7 Trajectory – movie (requires one PDB file with a series of models separated by the same tag)

Take a single trajectory and generate a series of images of contact maps that display both the contact map of the given frame, as well as a “fading trace” of contacts found in preceding frames. Very useful for tracing the evolution and fluctuation of contacts in simulations!

1.8 Contact map clustering (requires one PDB file with multiple models separated by the same tag)

Calculate the pairwise contact similarity for all models in a single ensemble file and plots it as a function of the model number. Clusters of models with contact similarity above a user-defined cutoff can be extracted and saved. It’s useful for clustering of models and identification of consensus structures, e.g. from *de novo* folding simulations.

1.9 Contact / distance map from file (requires one or two files compliant with either format mentioned below)

Reads and displays a contact map or a distance map saved in a file (ASCII type PHYLIP, CASP/EVA, CLANS or Microsoft Excel formats are currently supported – see example files). The user can also open two maps, combine them and save the result into a new file.

2. LOAD THE INPUT FILES

2.1 File format

PROTMAP2D accepts the following types of files:

• three-dimensional representation in Cartesian coordinates in the **PDB** format (Bernstein et al., 1977). In a file with more than one protein chain (a multimer for instance), different chains must bear different names (A, B, etc). Heteroatoms are excluded as non-aminoacids. It is recommended to avoid representing different residues with different types of atoms, e.g. some residues by all atoms and other residues by Cα atoms, because this would introduce bias in multi-atom metric (‘all-atom’ and ‘heavy-atom’) analyses.

• **Swiss-PdbViewer** (Guex and Peitsch, 1996) projects containing homology modeling layers as PDB models superimposed onto one another

• two-dimensional representation of contacts or distances as a matrix in the **PHYLIP** (Felsenstein, 1989), **CASP/EVA** (Grana et al., 2005) or **CLANS** (Frickey and Lupas, 2004) text formats

• two-dimensional representation of contacts or distances as a matrix encoded as a sheet in **MS Excel** 95+ format

2.2 Archived files

The user can also provide an archive of the file (a useful feature while analyzing huge model ensembles), in zip or gzip format, depending on the operating system. A tar archive compressed with zip or gzip is also acceptable.

Tarred or zipped files originating from PDB file sets should have no subdirectories. In other case the program reports file error. Also, each model has to terminate with the specific (common) tag, see 2.6. It is not recommended to assemble multiple files comprising more than one model per file, this feature is not supported.

2.3 Incomplete models (missing residues)

PROTMAP2D allows analysis of incomplete models. By default, the residues are indexed as they are in the input file (e.g. 10-100, 110-210 for a file with 10 residues missing in the N-terminus and 10 in the middle). Alternatively, the user can use the option to renumber the residues consecutively, from 1 to the total number of residues in the file. Beware that the latter option results in ‘compressing’ the gaps and should not be used if the user intends to compare models that differ with respect to the location and/or length of the gaps, as the corresponding residues may be renumbered differently!

In the case of multiple model analysis or model comparison incompleteness, the program reads the maximum residue index that serves as the final size of a map.

2.4 Missing atoms

Missing atoms are allowed, but the user should keep in mind that any inconsistencies of completeness within the model or between the models to be compares will influence the result. Thus, for protein models that partially comprise only C-alpha trace, and partially full-atom representation of residues, it may be advisable to avoid calculating maps using other metrics than C-alpha, ‘heavy’ or ‘any’ atoms, as otherwise C-alpha atoms may become invisible (e.g. if the C-beta metric is used). See: 3.1.1.

2.5 Sequence correspondence (compliance)

In order to allow the user to analyze any collection of protein structures, PROTMAP2D does NOT require the models to have identical sequences. In case of multiple models, the correspondence between residues (i.e. alignment) is derived from the numbering of residues in the input file.

CAUTION: If one attempts to compare homologous proteins, in which corresponding residues have different numbers (e.g. due to different insertions or terminal extensions), the program will NOT be able to produce a meaningful result. PROTMAP2D does NOT carry out alignment of contact maps, thus the user has to make sure the homologous residues to be compared MUST have the same numbers.

2.6 Multiple model files:

In ensembles or trajectories, different structures must be separated by a tag indicating the end of model (e.g. **ENDMDL**).

*CAUTION: the **TER** tag separates different chains within the same model, so for multimeric proteins it should never be used to separate different models.*

2.7 The preprocessing phase

PROTMAP2D scans the input file to identify the number and length of all polypeptide chains in the first model, total number of models in the file, the type of atoms available for analysis, and the availability of the secondary structure annotation. Depending on the result, different palettes of available options will be offered. To save the time, in the case of multiple model files (ensembles, trajectories – which can contain tens of thousands of structures!), only the first model in the file is analyzed, therefore the user must either provide the maximum length or make sure that all subsequent models do not contain residues with indices larger than the last residue of the first model, or that at least the number of residues is never greater than the number of residues in the first models, but in that case the residue renumbering option has to be turned on. Remember that in all trajectory-related jobs, PROTMAP2D numbers the first (supposedly native) structure as 0, not 1!

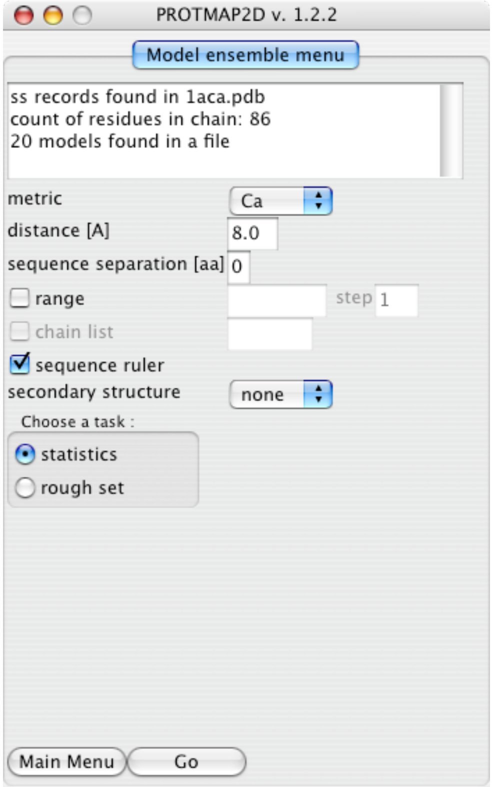
2.8 Contact / distance map from file

In this case, no 3D structure needs to be processed, so the user is not requested to provide any parameters.

3. PROVIDE JOB-SPECIFIC PARAMETERS

3.1 Common parameters

The user is asked to specify the following parameters (of which some are common and some unique to one or more jobs):



ss records found in 1aca.pdb
count of residues in chain: 86
20 models found in a file

metric: Ca
distance [A]: 8.0
sequence separation [aa]: 0
□ range   step 1
□ chain list
☑ sequence ruler
secondary structure: none

Choose a task :
● statistics
○ rough set

3.1.1 Metric

Depending on the input file (i.e. what kind of atoms are found), the user is asked to define the type of atoms to be considered for each residue to calculate the distance in Ångströms (with respect to the same type of atoms from all other residues):

- Ca: C-alpha atom
- Cb: C-beta atom (glycine has no C-beta, C-alpha atom will be used instead)
- heavy: all non-hydrogen atoms
- all: all atoms (including hydrogen)

3.1.2 Contact [Å]

Two residues are regarded as being in contact if the distance between their metric-specified atoms (see above) is equal or lower than this value. In case of 'heavy' or 'all' metrics, two amino acids will be considered in contact if any pair of the relevant atoms is found below the minimal distance.

3.1.3 Sequence separation [aa]

This parameter describes the minimal sequence separation between residues to be considered as in contact, i.e. it allows excluding contacts between nearest neighbors in the sequence. To exclude contacts between consecutive residues (e.g. residues 1 and 2), but retain the possibility of contacts between any other residues (e.g. residues 1 and 3), the value of this parameter should be set to 2.

3.1.4 Chain list

This option is used only if a model with multiple chains is provided (e.g. an oligomer or a complex, with different chains separated by the **TER** tag). It allows selecting only a subset of chains to be analyzed, by entering the comma-separated list of chain IDs. If a multi-chain file is analyzed, different chains will appear in different sections of the map, separated by white lines.

3.1.5 Sequence ruler

This option enables the display of the numbering of residues and secondary structure.

By default, PROTMAP2D uses the residue numbering used in the input files. In the case of multiple chains, however, residues of the second and subsequent chains are renumbered (the actual number of the residue in the Nth chain – first residue in the Nth chain + last residue in the preceding chain + 1) to avoid overlap as well as large gaps in numeration between the chains.

The user has also an option to renumber all consecutive residues starting from 1 for the first residue in the first chain, but this will ‘compress’ all gaps, if there are any.

3.1.6 Secondary structure

This option allows to either display the original assignment of helices and strands reported in the input file (if available, e.g. in the original files from the PDB) or to calculate them using third-party methods **DSSP** (Kabsch and Sander, 1983) or **STRIDE** (Frishman and Argos, 1995) programs. The availability of DSSP or STRIDE is checked by PROTMAP2D at the start-up.

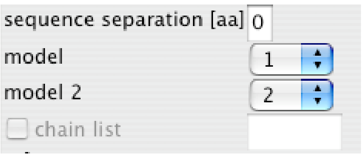
See *Appendix B* for the availability of third-party programs and installation instructions.

3.2 Model selection

Allows to select a particular model or a subset of models from a multiple model file.

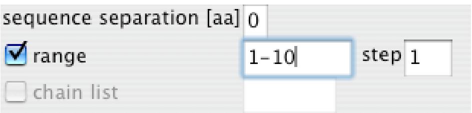
3.2.1 Model id selection

This applies to **single model** (‘model') and **model vs. model** ('model' and 'model 2') jobs. Model id is simply index numbers according to original model position in the input file.



sequence separation [aa] 0
model 1
model 2 2
□ chain list

3.2.2 Range and step for selection of models from trajectories and ensembles



sequence separation [aa] 0
☑ range 1-10 step 1
□ chain list

Trajectories from molecular dynamics may contain a large number of models. To avoid long calculations of all vs. all comparisons, PROTMAP2D offers the possibility to analyze only a subset of frames/models, indicated by the parameters “range”, e.g. 5000-10000 from the total of 10000, and “step”, which can be used to analyze e.g. every 10th model from the selected range. The default range includes all models and the default step is equal to 1. When comparing two trajectories, range and step can be defined independently for each of them.

NOTE: In all trajectory-related jobs, PROTMAP2D numbers the first (supposedly native) structure as 0, not 1, so the value of step set to 10 will result in comparing the native (0th) structure to models number 10, 20, 30 and so forth (until the end of the file or until the limit *specified by the range parameter). By default (i.e. if no range or step is indicated), all models will be compared.*

3.3 Tasks

Some of jobs' options dramatically change the resulting map and its semantics as well. That's why they are separated from others and called hereafter tasks.

3.3.1 Single Model

This job allows the user to compute either a **contact map** or a **distance map**. These two kinds of maps can be manipulated in different ways, using different options (See: 4.4 for details).

NOTE: Distance map calculation, even for single-atom metric (CA/CB) is dramatically time-consuming for proteins longer than 200 aa. For details see: 3.4.3 and 3.4.

3.3.2 Model ensemble

This job offers two kinds of statistical contact analysis. Pure contact occurrence frequency-based **statistics** or **rough set**- based (Pawlak, 1982). Rough set analysis is dedicated mainly to NMR ensembles while the output divides all the encountered contacts into two groups: “common” (present in each model) and “possible” (present in some models).

3.3.3 Trajectory

This job offers several alternative options (hereafter called tasks) to display particular features of multiple models:

3.3.3.1 Statistics

Trajectory - Statistics task performs the same analysis, as if the trajectory file was loaded as an ensemble. Here, the only difference is in the presence of the reference (native) structure in the trajectory (model number 0 by default or the first structure in the selected range), and the lack thereof in the ensemble. This function calculates the frequency of each contact in the whole ensemble/trajectory, and displays it using the grey scale (e.g. contacts occurring in 1-10% of models are very dark grey, contacts found in 41-50% of models are in medium grey, and contacts found in 91-99% or 100% of models are in white). The availability of the reference structure in the trajectory allows to display separately the native contacts, i.e. those that occur in the reference structure (lower left triangle) and the non-native contacts, i.e. those that do not occur in the reference structure (upper right triangle). The “Merge” button can be used to display all contacts in a single map. To undo this, click “Unmerge”.

3.3.3.2. History

Trajectory - History displays the ”memory” of contacts from the perspective of the last frame of the selected range. Here, the grey scale (up to 255 shades) is used to discriminate between the contacts that occurred only at the beginning of the trajectory and never appeared again (dark grey), from those that occurred late (light grey and white). This option provides a still picture, which is complementary to the output of the **trajectory – movie** job. ‘History’ differs in that if the number of models to be compared exceeds the length of 255, the set of models is divided into 255 sub-sets, from which the last model serves as the reference to derive the colors .

3.3.3.3 Contact persistence

Trajectory - Contact persistence identifies and displays contacts that, throughout the whole trajectory (or range), appeared in at least N (a number defined by the user) successive models. It can be used to discriminate between relatively stable contacts (e.g. salt bridges) and those that fluctuate, even if their overall frequency may be similar.

3.3.3.4 Contact number evolution

Trajectory - Contact number evolution generates a list of number of contacts within a user-defined region or between two user-defined regions, from the first model to the last. The region can be defined as a comma-delimited sequence of residues, containing single residues as well as 'from-to' intervals (e.g. 20,23,50-54 vs. 40-45,102-187). This option can be used to monitor the (un)folding of a particular region or interactions of two regions in the course of a simulation. The output is a text file with tab-delimited columns, which can be imported into any spreadsheet software to generate a plot.

3.4 Computing speed

Although most jobs are computed quickly, there are three exceptionally demanding job types that require special caution. It must be also mentioned that the selection of the distance metric may have an influence on computing time.

3.4.1 The usual case

Typical job and/or task (except cases mentioned below) calculation routine is founded on the *KDTree* algorithm (de Berg et al., 1997) as implemented in *BioPython* (Hamelryck and Manderick, 2003). This allows rapid calculation time, which increases in a manner close to a linear function of atom number x residue number x model number (see: 3.4.4). So even huge contact maps should be calculated within minutes on a modern workstation.

(A C-alpha, 8Å contact map for a 5000-residue protein is calculated in 7 minutes on a compact *Mac Mini 512MB RAM, 1,42 GHz PPC, OSX 4*. Trajectory statistics map for 430- residue 30- model ensemble calculates in about a minute.)

3.4.2 Speed choice

Special cases, **trajectory contact map comparison** and **contact map clustering**, require the program to store all the contact maps simultaneously in the RAM memory. Typical workstations with standard memory may not be able to process files that contain very large and/or numerous models. For this reason, we introduced a choice of "memory- safe" (default) or "speedy" options (one that assumes that all data can be stored in the memory without overloading it). The "memory safe" alternative should prevent the problems of overloading, but makes the computing time increases quadratically with the number of models.

(e.g. 2 s for 30- residue 25-model ensemble contact map speedy clustering against 40 s for the same result obtained in memory safe way, configuration as in 3.4.1).

3.4.3 Single model - Distance map

The distance map differs in the amount of information from the binary contact map, and so do the respective calculations with respect to the time of calculations. The distance map is in fact a two- dimensional graph of the distance function, and it requires explicit all-against-all distance calculations for all residue pairs. This may drastically slow down the calculation for large proteins. Calculation of a distance map for a protein of 200 aa requires approximately 100x more time than calculation of a contact map

(e.g. 0.5 s for 190- residue contact map against 50 s for distance map, configuration as in 3.4.1).

3.4.4 Speed and metric function

Multi-atom metrics ('heavy' or 'all') are more demanding than single atom metrics (CA/CB) simply because the program has to check all possible atom-to-atom distances for a pair of residues (e.g. about 400 possibilities instead of just one). Thus, calculation of distance maps based on multi-atom metrics may be more time-consuming for large proteins, compared to the single-atom metrics.

(e.g. 190- residue model contact map calculates respectively for CA/CB, heavy, all atoms: 1 s, 11 s and 12 s, configuration as in 3.4.1)

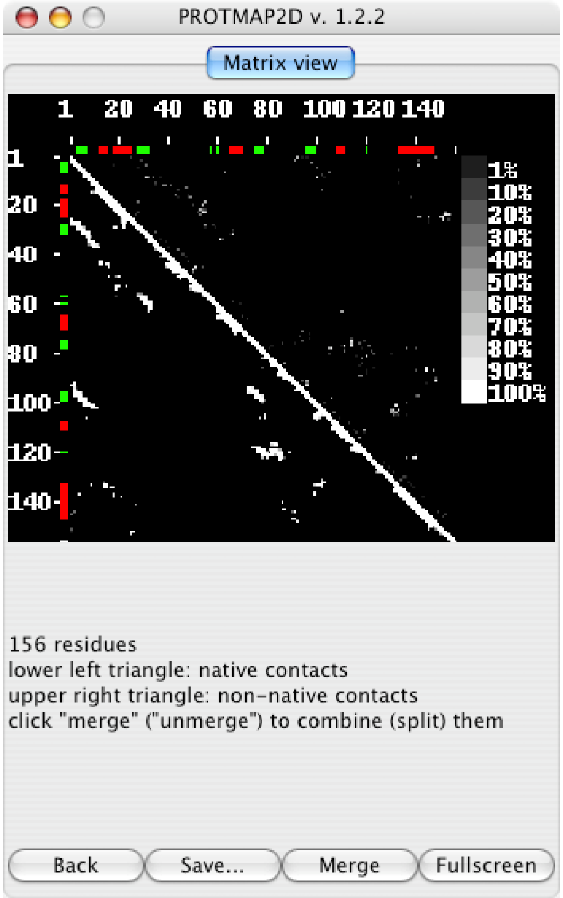
3.4.5 Progress of calculations

The progress bar is shown to inform the user about the relative time remaining for the calculation to finish. The user has a possibility to abort the job that turned out to be too demanding and return to the previous screen with options.

4. MATRIX VIEW

4.1 Map views

This is a graphical summary of all jobs (except **trajectory - contact number evolution**). In the case of contact maps, white and black dots indicate the presence and absence of contacts according to the specified criteria. Grey dots appear in analyses concerning multiple models and indicate contacts present in a subset of models, depending on the job. In the case of distance maps, the shades of grey indicate different distance ranges. If shades of grey are present in the picture, the grayscale legend is shown. Sequence ruler and/or secondary structure (red bars for helices, green for strands, white for gaps) may be are displayed depending on the options selection made earlier. If the analysis presents statistics for multiple models, then the secondary structure pattern is shown only for the reference model (first model in the ensemble or in the trajectory). In the case of individual frames of a trajectory, the secondary structure corresponds to the currently selected frame. If two models are compared, their respective secondary structures are shown in the upper and left part of the plot. As multiple shades are present in the picture, grayscale legend is included.



156 residues
lower left triangle: native contacts
upper right triangle: non-native contacts
click "merge" ("unmerge") to combine (split) them

By default, the map is scaled to the size of the program window. For most tasks the user can display the map in the resolution of one pixel per residue by clicking the ‘Full screen’ button. The user can also click on the map and draw a rectangle to select a section of the map that is displayed in a pop-up window, with an X/Y axis range indicator, which can be resized at will.

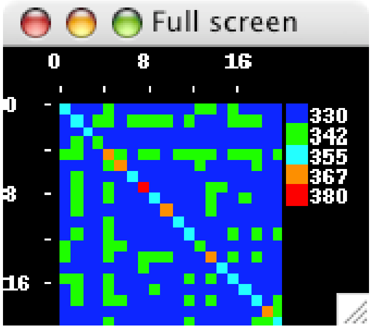
4.2 Color maps

4.2.1 Trajectory contact map comparison

In this case the output is a matrix, whose fields show, in a colored scale (see legend), the number of contacts that each contact map from one trajectory has in common with each contact map from the other trajectory. The rulers on the left and at the top show model indices from the 1st and 2nd trajectory, respectively.

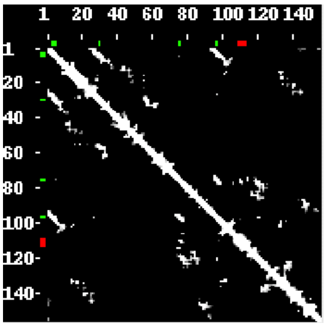
4.2.2 Contact map clustering

This job yields the same view. Here, the matrix is symmetrical and both ‘range’ and ‘range 2’ rulers are equal to the number of models in the file, making the matrix symmetrical. Additionally, **contact map clustering** allows selecting subsets of models whose contact maps are similar above the user-defined ‘cluster cutoff’. If this option is enabled, a new button ‘Save cluster’ appears, which can be used to save a list of numbers of members of the selected cluster as a text file.



NOTE: PROTMAP2D offers only a preliminary and naïve way of clustering protein conformations according to the similarity of their contact maps. However, the raw output of comparative analyses, indicating the values describing the % identity between contact maps can be saved as a matrix that may be further analyzed and clustered using specialized programs. The supported file formats include CLANS, which is an application dedicated to the clustering of protein data.

4.3 Movies



The result of the **trajectory - movie** job is an ordered series of images of contact maps (frames), which include the ‘memory’ of contacts from preceding frames, displayed as a trace fading to black in up to 254 shades of gray. The number of displayed ‘memorized’ frames (0 to 255 maximum) can be set using the 'fadeout' parameter. The frames can be scrolled back and forth manually, using the mouse roller (‘movie’), or presented as a bidirectional 'animation'. Animation can be exported into MPEG file.

4.4 Distance / Contact Map From File special features

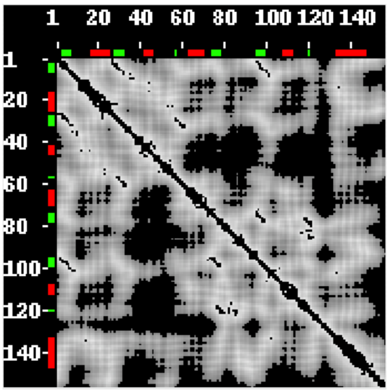
4.4.1 Dual mode: Distance/Contact or Fuzzy/Contact

The prime mode shows the original result, i.e. a distance map or a fuzzy contact map read from the file(s) (e.g. ensemble – statistics result), while the secondary mode shows the binary (pure contact map) version of the original map, calculated according to the defined metric. There's a special button for mode switching between the two modes of display.

4.4.2 Floating cutoff: distance or probability

This option lets the user increase/decrease the contact threshold using the mouse roller, similar to **trajectory movie** (see: 4.3) and observe its effect on the map in the real time.

4.4.3 Upper/Lower/Sequential threshold



The user, knowing the maximal value shown on the map (minimum is zero by default), is allowed to define independent thresholds and thereby turn off the display of points with values below or above the thresholds.

This feature allows to visualize contacts (or other values specified by the matrix, e.g. contact probability) only within the desired range of values. The map, in any stage of the manipulations described above, can be saved “as is”.

4.5 Saving the results

4.5.1 The usual case

In the matrix view, there is always a 'Save...' button that gives opportunity to save the graphics in a bmp, gif, jpeg, png, or tiff format.

It is also possible to save the results as a PHYLIP, CASP, EVA, or CLANS ASCII file, or a MS EXCEL-formatted spreadsheet comprising the distance matrix with supplementary information. See *Appendix A* for detailed description of file formats.

4.5.2 Trajectory – movie

In this case, apart from all default output capabilities, the user is also able to save a movie record as an animation mpeg file. This feature is not yet supported under MacOSX and will be added in the future. (See also: 4.3).

NOTE: The view must be in a "still" mode (so "Movie" or "Animation" must not be running) in order to save the movie.

4.5.3 Trajectory / contact number evolution

This job does not produce any map, only the text output. All graphics saving options are disabled.

5. Another analysis

Selecting 'Back' returns the user to the previous screen without saving the results (i.e. all calculations and maps are lost). However, the input files remain pre-processed, and the program should remember the previously provided job-specific parameters, facilitating recalculation of results.

Credits

Michał Piętal designed and implemented the algorithm. Irina Tuszyńska tested PROTMAP2D performance and provided many useful suggestions and comments. Janusz Bujnicki conceived of the program and participated in the design and coordination of the study. This project was supported by the Polish Ministry of Scientific Research and. Information Technology (grant PBZ-KBN-088/P04/2003 for the development of software tools for bioinformatics, genomics, and proteomics). Michał Piętal was supported by the NIH (Fogarty International Center grant R03 TW007163-01).

Literature References

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer-EE, J., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) The Protein Data Bank: A computer-based archival file for macromolecular structures. J.Mol.Biol., 112, 535-542.

Guex, N., Peitsch, M.C. (1996) Swiss-PdbViewer: A Fast and Easy-to-use PDB Viewer for Macintosh and PC. Protein Data Bank Quaterly Newsletter 77, pp. 7.

Felsenstein, J. (1989) PHYLIP - Phylogeny inference package (version 3.2). Cladistics, 5, 164-166.

Grana, O., Baker, D., Maccallum, R.M., Meiler, J., Punta, M., Rost, B., Tress, M.L., Valencia, A. (2005) CASP6 assessment of contact prediction. Proteins,

Frickey, T. and Lupas, A. (2004) CLANS: a Java application for visualizing protein families based on pairwise similarity. Bioinformatics, 20, 3702-4.

Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, 22, 2577-2637.

Frishman, D. and Argos, P. (1995) Knowledge-based protein secondary structure assignment. Proteins, 23, 566-579.

Pawlak, Z. (1982) Rough sets. Internat. J. Comput. Inform, 11:341-356

de Berg, M., van Kreveld, M., Overmars, M., Schwarzkopf, O. (1997) Computational Geometry-Algorithms and Applications. Springer-Verlag

Hamelryck, T. and Manderick, B. (2003) PDB file parser and structure class implemented in Python. Bioinformatics, 19(17) 2308-10

Appendix A.

CONTACT MAP TEXT FILE FORMATS

A1.1 Phylip

The text output includes the matrix in the format compliant to that used in the **PHYLIP** package to represent evolutionary distances between species (Felsenstein, 1989).

The beginning of the file is a number that is, length of the protein as an integer.

In order to describe the sequence information first 11 characters corresponding to the 'species' name are used by PROTMAP2D. In other matrices, (e.g. results of **trajectory contact map comparison** and **contact map clustering**) all these positions are blank; only the model number from first set is listed as row id.

position	value
1	chain id
2 - 5	absolute residue id
6 - 8	residue name
9	space (blank)
10	secondary structure: H (helix) / E (strand) / - (other) / blank if not chosen
11	space (blank)
12 - ...	matrix records, as formatted numbers, delimited by space (see also: A3)

A1.2 CASP

This format is used in a **CASP** experiment since the **Contact Map Prediction** (**RR**) category was introduced (Grana et al., 2005).

The file consists of four sections:

- **PFRMAT RR** fixed contact map indicator tag
- **REMARK** this part is used to store additional information (see: A2)
- **SEQRES** single-letter protein sequence (and thus non- compliant with PDB (Bernstein et al., 1977) but compliant with the CASP format spec.)
- contacts part

Contact map is represented as a non-symmetric non-trivial (excluding contacts of a residue with itself) list of pairs been in contact (or having positive contact probability/ frequency), sorted by the contact value (see: A3)

position	value
1 - 5	residue1 id
6	space (blank)
7 - 11	residue2 id
12	space (blank)
13 - 17	'0' character (to uphold compliance)
18	space (blank)
19 - 22	contact distance (in A)
23	space (blank)
24 - 28	contact value (see: A3)

A1.3 EVA / EVAcon

EVA is a continuous benchmarking experiment independent of CASP (Grana et al., 2005). **EVAcon** deals with measures regarding contacts, including contact map prediction.

EVA format is very similar to CASP, (see A 1.2) their contact map representation differs only slightly. Introducing the **CONTC** tag makes EVA format PDB-like (record-based, tag-semantics). Another difference is that a sequential information (amino acid codes) is included in the contact record.

position	value
1 - 5	**CONTC** tag
6	space (blank)
7 - 11	residue1 id
12	space (blank)
13 - 17	residue2 id
18 - 22	(blank)
23	aa1 single-letter code
24 - 28	space (blank)
29	aa2 single-letter code
30 - 34	space (blank)
35	'0' character (to uphold compliance)
36	space (blank)
37 - 40	contact distance (in A)
41	space (blank)
42 - 46	contact value (see: A3)

A1.4 CLANS

CLANS (Frickey and Lupas, 2004) is a JAVA-based program for clustering, mainly for sequence comparisons, where the distances usually represent sequence similarity between individual proteins, not a physical distance between their residues. CLANS can be, however, used to analyze any type of distance matrices and we used it for clustering of proteins according to the similarity of their contact maps. Thus, CLANS itself and its data format is useful for post-processing of results from PROTMAP2D jobs such as **trajectory contact map comparison** and **contact map clustering**.

CLANS has numerous input/output formats, and the one used by PROTMAP2D and described here is called Attraction Matrix. The file consists of the following consecutive sections:

• **sequences=**N	N is the protein length
• #...	summary (see: A2)
• **<seqs>**	fixed: sequence section start
• **>**sequence_names	this section is constructed exactly as in the Phylip case (see: A 1.1)
• **</seqs>**	fixed: sequence section end
• **<mtx>**	fixed: matrix section start
• matrix	formatted as in the PHYLIP format (characters above 12, see: A 1.1)
• **</mtx>**	fixed: matrix section end

A1.5 MS Excel

Format of the **MS Excel** file output is binary, thus the description here concerns file display in any of MS Excel 95+ (*.xls) file readers (*MS Excel*, *OpenOffice Calc*, *NeoOffice* etc.).

Contact map is encoded in a worksheet named “Contact Map”, very similar to the PHYLIP format (See: A1.1). The exception is that instead of the value describing protein length in the first row, residue id information is placed in subsequent columns, in the same ‘1-11 characters‘ row format. Both residue information and the matrix values are placed separately in worksheet cells.

A.2 Summary and statistics

Besides the matrix itself, the text output includes the summary:

- number of models
- distance metric
- sequence separation (if enabled)
- secondary structure source (if enabled)
- number of residues analyzed
- chain filter (if enabled)

For each map, the following contact statistics are displayed:

- number of the native / non-native / all contacts
- percent ratio of the native / non-native / all contacts, as compared to the native structure
- percent ratio of secondary structure (if secondary structure was enabled)

Summary and statistics placement in the file may vary, depending on the file format.

Table I; Summary and statistics in files of different format.

format	special mark	placement against matrix	placement in file
Phylip	-	below matrix	at the end
Clans	**#** (comment)	above matrix	after length declaration
CASP / EVA	**REMARK** tag	above matrix	after filetype indication tag
Excel	-	separate worksheets	‘Summary’ and ‘Statistics’

A3. Numeric values

All values in the matrix/contact are formatted as four-digit numbers, that is three-digit precision plus a decimal point (except for distance and numeric jobs, where the number of digits may vary, see: Table II), and are separated by a single space in ASCII matrices (PHYLIP, CLANS) or placed in different cells (EXCEL).

Table II; Semantics of contact values.

matrix kind	contact values	semantics	representing jobs/tasks
contact map	0.000 and 1.000	1.000 = contact	contact map contact persistence
comparative	0.000, 0.500 and 1.000	1.000 = common contact 0.500 = a contact	model vs. model rough set
statistical	0.000 – 1.000	contact frequency given a model set	statistics
distance	0.00 – max distance	distance real value	distance map
numeric	0 – max number	common contact number for a model pair	contact map comparison contact map clustering

Appendix B.

EXTERNAL (THIRD- PARTY) PROGRAMS

B1. Linux installation prerequisites

In the case of Linux systems, due to some technical limitations, all packages used by the program that are not explicitly programmed by the authors of PROTMAP2D, must be installed on the system, BEFORE the installation of PROTMAP2D:

Table III; Required package list and versions

name	Debian (*deb*) package name	minimal&recommended version
Python 2.4	python2.4	2.4.2
BioPython (full)	python2.4-biopython	1.42
PIL	python2.4-imaging	1.1.5
wxPython	python-wxgtk2.6	2.6.1
PyExcelerator	-	0.6.3a
PyMedia	-	1.3.5

*NOTE: **PyExcelerator** and **PyMedia** packages can be installed by downloading from their sites: http://sourceforge.net/projects/pyexcelerator and http://pymedia.org, respectively.*

B.2. Secondary structure programs

To enable calculation and visualization of secondary structures, the user is allowed to use third-party programs **DSSP** (Kabsch and Sander, 1983) and/or **STRIDE** (Frishman and Argos, 1995).

Third-party programs are NOT included in the PROTMAP2D package, however the user is encouraged to download and install them, rename (make sure the name of the executable file is lower-case, and that the Windows version contains the “.exe” extension) and put the binaries into the PROTMAP2D's directory or make them visible at the defined path visible for the operating system.

The following web pages contain download instructions: http://swift.cmbi.ru.nl/gv/dssp/ (dssp) and http://wolf.bi.umist.ac.uk/unix/stride.html (stride). See the Table IV for Support of Operating Systems for Secondary Structure assignment programs (below).

Program	MacOSX	Windows	Linux
DSSP	+	+	+
Stride	+	-	+