RNAmap2D v. 1.7.5

User Manual

RNAmap2D is an application for the analysis of RNA structures and structures of protein-RNA complexes through two-dimensional contact maps and distance maps.

To obtain a contact map with **RNAmap2D**, 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 way in which residue positions are extracted.
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 to an output file.

Table of Contents

1. Types of jobs

Single model (requires one PDB file)

Calculates and visualizes a contact map or a distance map for one RNA structure including ions and ligands, additionally pairing types and stacking classes. The map shows dots where two residues are in contact, and they can be colored according to base pairing types, base stacking, and several other types of contacts. Any PDB-compliant structure containing RNA should be suitable for this.

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

Analyzes the frequency of different contacts in a set (ensemble) of alternative models, e.g. an NMR ensemble or a cluster of decoys obtained from *de novo* folding analysis. It is required that the PDB file contains multiple models having identical sequence and length.

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

Compares contacts in two different models of the same RNA structures, e.g. a theoretical prediction vs. the experimentally determined structure or models obtained under different conditions, two close homologues etc. This requires one PDB ensemble containing at least two models or two PDB files with different models. The models don't have to be of the same length, but in both models in question, chain identifiers must be the same, for multimeric structures. Comparing two ambiguous multimeric structures of different length may be sometimes erroneous.

1.4 Contact map from file (requires a contact map encoded in CSV, CASP/EVA, Clans, Phylip or Excel format, a single file or two files)

Reads and displays a contact map (binary or fuzzy) or a distance map saved by RNAmap2D or a different program in a specific file format (ASCII type PHYLIP, CSV, CASP/EVA, CLANS or Microsoft Excel formats are supported). The user can also open two maps, combine them and save the result into a new file.

1.5 Complex contact map (requires one PDB with RNA and protein chains as a complex)

Calculates and visualizes a contact map for a structure of a protein-RNA complex. This requires a PDB file with both RNA and protein chains present.

1.6 Complex docking (requires PDB file with multiple models of RNA and protein as a complex)

Calculates and visualizes a docking model set. The analysis can output a statistics map, showing contact frequency against reference model or contact evolution throughout the models set, given specified residue range.

1.7 Secondary structure (requires an RNA secondary structure file in VIENNA, BPSEQ or CT format)

Reads and displays a contact map given a file with RNA secondary structure file (VIENNA, BPSEQ, CT formats are supported). The map can then be saved in a form of a contact map, either as a picture or text representation.

2. Input file formats and processing

2.1 File formats

RNAmap2D accepts the following file types:

- A 3D structure representation in the **PDB** format (Bernstein et al., 1977). Residues are numbered as they appear in the file (exceptions: complexes, see notice in 3.3 , but also ligands/ions). Ligands and ions are marked by the parser as an artificial chain '!'. It is recommended that representing different residues with different types of atoms should be avoided, e.g. some residues by all atoms and other residues by C1' atoms, because this would introduce a bias analyses, that is, sometimes lack of a contact.

- A pair of PDB structures that are compared. It is recommended that the residue identifiers of both structures be identical in both files. A multi-model PDB file. Each model should terminate with the **ENDMDL** tag, if any other terminal tag is present in the file, the user is given an option to input this tag, in order for the program to read out all models as separate entities.

- Two-dimensional representation of contacts or distances as a matrix in the **CSV** (coma separated values), **PHYLIP** (Felsenstein, 1989), **CASP/EVA** (Grana et al., 2005) or **CLANS** (Frickey and Lupas, 2004) text formats (see Appendix B for details).

- Two-dimensional representation of contacts or distances as a matrix encoded as a sheet in **MS Excel 95+** (BIFF) format (see Appendix B for details).

- RNA secondary structure text files of the following popular formats: **Vienna**, **BPSeq** and **CT**.

2.2 Archived files

The user can also provide an archive of the input file (useful for analyzing i.e. very large model ensembles), in the .zip, .gz, .tar.zip, or .tar.gz/.tgz format, depending on the operating system (.tar and .gzip files can't be read on Windows). Tarred or zipped files originating from PDB file sets should have no subdirectories. In other cases the program reports a file error.

2.3 Missing residues and atoms

RNAmap2D allows the analysis of incomplete models. By default, the residues are indexed as they are in the structure file, but with no gaps (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 'structure-based' option to renumber the residues consecutively, from 1 to the total number of residues in the file. Third numbering option is 'residue id-based (gapped) which causes inserting gaps into the map, if structural gap is encountered in a PDB file. This third numbering option is not present in complex docking job.

However beware that the 'structure-based' or 'resid-based (ungapped)' numbering options will result 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. In such cases, the corresponding residues may be renumbered (and thus placed on a map) 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. Except for 'residue id-based (gapped)' numbering scheme, the two remaining schemes result in the same contact map. The only difference is that in 'structure-based' case, all residues are numbered consecutively, starting from 1, regardless from the residue PDB numeration. 'Residue id-based (ungapped)' differs in that residue numbers are preserved as originating from the PDB file (i.e. for multimeric structures there can be multiple residues that have id equal to 1, 2 etc.).

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 compared will influence the result (in particuler, if any missing atom is searched for, because of the metrics chosen).

2.4 Sequence

When comparing two structures, in order to allow the user to analyze any collection of RNA structures, RNAmap2D does NOT require the models to have identical sequences. In case of multiple models, the correspondence between nucleotides (i.e. alignment) is derived from the numbering of nucleotides in the input file, but also depending on selected numbering scheme. However, when comparing two multimeric structures, chain identifiers must match because the program compares only those chain that occur in both models.

CAUTION: If one attempts to compare two homologous RNA structures, in which corresponding nucleotides *have different numbers, the program will NOT be able to produce a meaningful result. RNAmap2D does NOT carry out an alignment of contact maps, thus the user has to make sure the homologous* nucleotides *to be compared have the same numbers.*

2.5 Atom-based vs. residue-based definition

RNAmap2D can handle PDB files containing models of nucleic acids, including ligand and ion molecules or protein chains. The user is given an option to choose one of several metric functions and define contact threshold in order to obtain custom contact map.

Please note that contact map or distance map is a mapping made on the sequence level (1D sequence onto 2D map). Thus a residue, either nucleotide, amino acid or ligand/ion, becomes the respective matrix index. Contact maps and distance maps provide uniform information regardless molecule length, size or nature (also mixed protein-RNA or RNA-ligand maps are properly defined).

Each residue is built from a variable number of atoms which vary depending on residue class (nucleotide, amino acid, heteroatom) but also type (i.e. glycine and tryptophan in proteins). Thus our program RNAmap2D, just like its predecessor PROTMAP2D, defines map index on residue level. This approach is confirmed in case of many other tools, works and analyses, however it is theoretically possible to define a correspondence where each atom in a molecule is mapped into unique contact or distance map index. This approach however might cause difficulties in case of i.e. residue natural atom numbering conflicts, atom modifications or deletions, PDB files inconsistencies and errors etc.

This is why we define contact and distance maps as matrices mapping residue pairs on residue proximity features.

2.6 Preprocessing phase

RNAmap2D scans the model first to identify:

- the number and length of all nucleotide chains (first model)
- the number and type of ligands and ions (first model)
- discontinuous chains and the number of interruptions in the chains (first model)
- the number of modified base pairs (first model)
- total number of models in the file
- protein secondary structure (for complex job only).

Depending on the result and job type chosen, different palettes of available options will be offered. To save time, in case of multiple model (ensembles), only the first model in the file is analyzed in details, therefore the user must make sure that all subsequent models do not contain nucleotides with indices larger than the last nucleotide of the first model, or that at least that the number of nucleotides is never greater than the number of nucleotides in the first model. In such case, the nucleotide renumbering option has to be switched on.

Note that in case of contact map from file and secondary structure jobs, no preprocessing or job specific parameters are given.

3. Job-specific parameters

3.1 Common parameters

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

3.1.1 Metric

The user is asked to choose the type of atoms to be considered to calculate the distance in Ångströms.

3.1.1.1 RNA structures

For an RNA structure the following metric options are available:

- C1'
- C4'
- O5'
- N1/N9 (for purines and pyrimidines, respectively)
- heavy: all non-hydrogen atoms
- all: all atoms (including hydrogen)

3.1.1.2 Proteins

For protein chains following metric options are available:

- 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)

3.1.1.3 Ligands and ions

In case of single model job, ions and ligands are extracted from **HETATM** PDB records. Additionally program searches for uncommon residue names in ordinary ATOM records, compared against internal nucleotide modifications, ligands and ions lists. After excluding water atoms, all other atoms are collected from the molecule and represented as the last pseudochain in the contact map, also possible to produce contacts with RNA basepairs..

3.1.2 Contact distance [Å]

Two residues are considered 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 nucleotides will be considered in contact if any pair of the relevant atoms is found below or at the minimal distance.

3.1.3 Sequence separation [nucleotides]

This parameter describes the minimal sequence separation between nucleotides to be considered as in contact, i.e. it allows excluding contacts between the nearest neighbors in the sequence. To exclude contacts between consecutive nucleotides (e.g. nucleotides 1 and 2), but retain the possibility of contacts between any other nucleotides (e.g. nucleotides 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 (different chains separated by the **TER** PDB tag), and for certain jobs. 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 (if this option is switched on, see: 3.1.6).

3.1.5 Sequence ruler

This check box enables the display of the numbering of nucleotides.

By default, RNAmap2D uses ‘residue id-based (ungapped)’ numbering used in the input files (except for the complex docking job). In such case, residue numbering on the ruler will be retained as originally found in the PDB file.Similar numbering scheme will be also present in case of ‘residue id-based (gapped)’ numbering, but in that case, N-terminal gaps as well as gaps within a chain can be introduced, depending on the contents of a PDB file. In case of ‘structure based‘ numbering, the expected residue numbering to be found in the ruler are simply natural numbers, starting from 1 and ending at the number of residues in the model.

Additional feature of the ruler is to visualize RNA secondary structure. The structure is calculated by by either using RNAVIEW plugin or internal algorithm, when RNAVIEW is not present. Secondary structure is coded as pink and violet sequential blocks, resembling opening and closing brackets in Vienna sequential format, respectively.

Because the internal routine calculating secondary structure (pairings) can be slow for large RNAs, there is an additional option to skip this routine and obtain a contact map very quickly, in case of ‘single model’ job. To do so, the user is advised to select ‘contact map no pairings (fast)’ task in single model job. In all other cases, secondary structure calculation is enabled, unless the total residue number exceeds 300 and RNAVIEW plugin is missing.Also in case of complex PDB files, in ‘complex contact map‘ job it is possible to visualize protein secondary structure. If the said PDB file contains secondary structure records (i.e. **HELIX**, **SHEET** or **TURN** tags), RNAmap2D will visualize it on the ruler, above the protein part of the map, while red bars symbolize helices, green symbolize sheets and. White bars visualize missing residues (structural gaps).

3.1.6 Chain boundaries

For jobs supporting chains selection, it is possible to toggle the visualization of chain boundaries in the final view. Borders are marked as white horizontal and vertical lines.

In case of complex jobs, the boundary is only marked between protein and RNA molecule, for better picture readibility and quality.

3.2 Model selection

Allows selecting 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 ensemble, complex contact map** ('model'), **model vs. model** ('model' and 'model 2') jobs. Model id is simply an index number according to original model position in the input file, starting from 1.

3.2.1 Range and step for selection of models from ensembles

This applies to **model ensemble** and **complex docking** ('range') jobs. Ensembles may contain a large number of models. To avoid long calculations, RNAmap2D 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 ("step 10") model from the selected range. The default range includes all models and the default step is equal to 1.

3.3 Residue id selection

This parameter applies exclusively to **complex docking** ('range') job. Contact number evolution is designed specifically for in-depth analysis of contacts of protein-RNA interfaces. The user can specify an ambiguous residue set (R1), also another independent set can be defined (R2), together forming an interface. Then, instead of contact map calculation, only contacts between R1 and R2 residues are recorded and output as a chart (see also: 3.4.3).

*Important notice: In case of a **complex docking** job (but also in the case of **complex contact map**), numbering of residues might vary from that originating from the file. This is because RNAmap2D collects all protein chains to place them before RNA chains in the contact map. This is to achieve appealing visualization of contacts, where it is easy to analyze inter- and intra-molecular contacts that are better distinguishable in the picture.*

As an example, if the preprocessing report discloses count of residues in chains (RNA) A: 49, count of residues in chains (protein) b:149, c:149, one should consider the following numeration to be found on the final map:

1 – 149	*protein chain B*
150-298	*protein chain C*
299-347	*RNA chain A*

Please note that in case of complex docking job, 'resid-based (ungapped)' numbering is disabled not to cause the unambiguity of residue identifiers.

3.4 Tasks

Options of some jobs change dramatically the resulting map and its semantics as well. That is why they are separated from one another other and called tasks.

3.4.1 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: 4.2.5.1 and 4.2.4 respectively, for details). In this job, additional **contact map no pairings** task is introduced for saving calculation time if no RNAVIEW plugin is present, no pairing information is important and the RNA molecule is large enough to cause calculation delay (see also 3.5.2).

NOTE: Distance map calculation, even for single-atom metric (e.g. C1') is very time-consuming for RNA longer than 100 nucleotides. For details see: 3.5.3 and 3.5.4.

3.4.2 Model ensemble

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

3.4.3 Complex docking

This job offers also two kinds of statistical contact analysis. Pure contact occurrence frequency-based **statistics** or **contact number evolution**. In the case of statistics, reference contact map is calculated, basing on first model serving as reference model (in the reported statistics, first model has id of 0th, the rest of models is shifted likewise). Then, all other contact maps are compared to the reference contacts, simultaneously collecting the frequency of contact occurrence. At the end, statistical map is similar to this of **model ensemble statistics** but all contacts can fall into two distinct groups: those existing originally in the reference contact map (lower-left triangle on a map), or those that were not present in the reference map (upper-right triangle).

Contact number evolution is designed specifically for in-depth analysis of contacts of protein-RNA interfaces. The user can specify an ambiguous residue set (R1), also another independent set can be defined (R2), together forming an interface. Then, instead of contact map calculation, only contacts between R1 and R2 residues are recorded and output as a chart. As in the previous case, there is a separate counter between contacts coexisting in a reference model and the other ones. See also: a notice in 3.3.

3.5 Speed issues

Although most jobs are computed quickly, there are more demanding job types that require special caution or other specific and exceptional issues. It must be mentioned that the selection of the distance metric may have an influence on computing time.

3.5.1 The usual case

The typical job calculation routine is based on the *KDTree* algorithm (de Berg et al., 1997) as implemented in *BioPython* (Hamelryck and Manderick, 2003). This allows rapid calculation, where time increases in a close to linear function of *(atom number x residue number x model number)*. Thus, even huge contact maps should be calculated within seconds on a modern workstation.

3.5.2 Single model – contact map

RNAVIEW plugin has quick algorithms that calculate pairings which are typical to single model contact map job. Unfortunately RNAVIEW to date is not available to all three platforms. This is why we designed our own code that performs pairing calculation when no RNAVIEW is present. Hence it is a rare case, but for large RNA structures it takes fast to calculate a contact map alone, but it might take much longer to calculate nucleotide pairing list and types. So this is the only case when calculations might be significantly slower than expected from contact calculation component alone. This is why, pairing calculation was excluded for extraordinary large molecules having 300 or more bases.

We highly recommend that the user ought to install RNAVIEW program as plugin used by RNAmap2D, on its supported platforms. Additionally, for such cases we designed **contact map no pairings** task which disables pairing calculations, and results in rapid contact map calculation for all cases.

3.5.3 Single model - distance map

The distance map differs in the amount of information from a 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 a large RNA.

3.5.4 Speed and metric function

Multi-atom metrics (‘heavy’ or ‘all’) are more demanding than single atom metrics (i.e. C1′ for RNA or CA/CB for proteins), simply because the program has to check all possible atom-to-atom distances (a pessimistic case) for each pair of residues (e.g. over 100 possibilities instead of just one). Thus, calculation of distance maps based on multi-atom metrics may be more time-consuming for large RNA and complexes, compared to the single-atom metrics.

3.5.5 Speed in text jobs

In all jobs reading out text files, the calculation speed is rapid since the only task for the program is reading contacts/distances and generating a picture.

4. Matrix view

4.1 Types of maps

Maps are the graphical result of all jobs (except for **complex docking 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 appearing in **model ensemble rough set** indicate contacts present in a subset of models, depending on the job. In case of **secondary structure** job, grey dots mean that the contact forms an RNA pseudoknot. In **model vs. model** case, grey dot symbolizes a contact which occurred only in one model among two been analyzed.

When there are multiple shades of grey present (**model ensemble statistics, complex docking statistics** and possibly fuzzy **contact map from file**), the shade intensiveness is frequency/probability visualization whilst black is a 0.00 value, white is a 1.00 value. The detailed values are possible to be encoded in any text (i.e. CSV) or MS Excel output format.

In the case of **single model distance map**, the shades of grey indicate different distance values: white color indicates distance equal to zero Ångströms, black contact symbolizes the residue pairs which distance is maximal among all distances in the structure. The sequence ruler and/or chains borders are displayed depending on the options selection made earlier. This also applies to distance map read out from file.

4.2 View options

4.2.1 Full screen and zoom pop-up

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 3x3 pixels per one contact, by clicking the ‘1con=9px’ 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 might be resized at will.

Important Notice: To enable pop-up rectangle selection, OSX users should press left mouse button while hovering mouse cursor over a contact map. This is because in OSX system an application window looses focus while mouse cursor leaves the application window area.

4.2.2 Spying glass

When the user points the mouse cursor over the computed map, the INFO window will be displayed. It displays information about the highlit contact: residue names (resi), indexes and information if the two residues are in the contact (contact: Y) or not (contact: N). Note that this applies to nucleic acid residues only.

Important Notice: To enable spying glass window refreshing, OSX users should press left mouse button while hovering mouse cursor over a contact map. This is because in OSX system an application window looses focus while mouse cursor leaves the application window area.

4.2.3 Merging and Averaging

For maps that aren't symmetrical (e.g. in complex docking – statistics case, but also for maps read out from two files), special buttons appear. If two distinct maps (i.e. left-lower versus right-upper part) have no common entries, 'Merge' button appears which makes the map symmetrical. In case if the map isn't symmetrical, similar button pair appears for a change, that is 'Average'. This time, all corresponding values are added and divided by two to make a map which is symmetrical and is the average of the two maps originally shown in different sections of the the picture. Those two operations can be undone by their twin buttons ('Unmerge' and 'Unaverage', respectively)

4.2.4 Distance Map / Contact / Distance map from File – special features

4.2.4.1 Dual mode: Distance/Contact or Fuzzy/Contact

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

4.2.4.2 Floating cutoff: distance or probability

This option, present in contact mode, lets the user increase/decrease the contact threshold using the mouse roller and observe its effect on the map in the real time. *Note that this process is gradual and slow for larger molecules.*

Important Notice: To enable contact map refreshing by change of floating distance/probability threshold, OSX users should press left mouse button while hovering mouse cursor over a contact map. This is because in OSX systems an application window looses focus while mouse cursor leaves the application window area.

4.2.4.3 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. *Note that this process is gradual and slow for larger molecules.*

4.2.5 Option panel: colors

This panel is accessible for the **single model – contact map** and for the **complex contact map** jobs. Options included in this panel vary depending on the selected job.

4.2.5.1 Single model – contact map

The option panel allows coloring specific types of contacts observed in RNA structures if they are identified by the program. The option panel also allows coloring identified ligands and ions and to visualize stacking. See also 4.3.1.2.

Additionally, other (not else classified) contacts can be given a color as well as the picture backround, for obtaining convenient pictures for the purpose their inclusion in scientific papers, websites or presentations.

NOTE: Pairing calculation using RNAmap2D internal algorithm is time-demanding as the calculation time might grow significantly for larger molecules. This is why, the authors decided to disable this feaure for molecules of the length exceeding 300 nucleotides.

4.2.5.2 Complex contact map

The option panel allows assigning colors to contacts within RNA, within protein and between RNA and protein, also to change background color.

4.3 Saving the results

4.3.1 Saving a contact map

4.3.1.1 Graphics

In the matrix view, there is always a 'Save...' button that gives opportunity to save the graphical representation of the map as a .bmp, .gif, .jpeg, .png, or .tiff file.

Also in case of zooming a map fragment pop-up window appears, as described in section 4.2.1. To save the fragment as a picture, the user should choose s ‘Save…’ option from menu.

4.3.1.2 Text

It is also possible to save the results as a PHYLIP, CASP, EVA, CLANS, MS Excel (BIFF) or CSV file. If the user uses the coloring option, colored contacts will be saved to a file as well and can be read out and displayed as originals (please note that this is a beta feature).

4.3.2 Saving the secondary structure

In the two following cases: **single model contact map** and **secondary structure**, the user is given the additional possibility to save secondary structure in the Vienna format. In the first case, only specific contacts are coded as secondary structure elements (determined as base pairs by using either RNAVIEW program or RNAmap2D internal routine), in the latter case, the whole map is saved because it always represents secondary structure. If RNA pseudo-knots are determined, they are encoded in Vienna file using special Vienna notation.

Please note that in single model job, secondary structure encoding requires that calculation of base pairs’ types be enabled.

Appendix A – Installation requirements

A1. Linux

A1.1 Installation instructions under most recent and stable Ubuntu Linux (12.4 LTS)

Below there are step-wise instructions of installing and running RNAmap2D on a blank Ubuntu 12.4 system. Please note that python2.7 distribution version as well as libraries are used. The instructions are a copy of the recipe to be found in README file, placed in RNAmap2D distribution folder.

1. Installing all available APT packages

```
sudo apt-get install python2.7-dev
sudo apt-get install python-wxgtk2.8
sudo apt-get install python-numpy
sudo apt-get install python-egenix-mxtexttools
sudo apt-get install python-excelerator
sudo apt-get install python-cogent
sudo apt-get install g++
```

2. Installing the remaining packages from source

a) Numeric

- download Numeric-24.2.tar.gz file from http://sourceforge.net/projects/numpy/files/Old%20Numeric/24.2/
- change to the directory with the file
- run following steps:

```
tar -zvxf Numeric-24.2.tar.gz
cd Numeric-24.2/
python setup.py build
sudo python setup.py install
```

b) BioPython

- download biopython-1.42.tar.gz file from http://biopython.org/wiki/Download
- change to the directory with the file
- run following steps:

```
tar -zvxf biopython-1.42.tar.gz
cd biopython-1.42/
cp setup.py setup.py.old
```

- edit the setup.py file, so that the diff command provides following output

```
diff setup.py.old setup.py
387c387
< #   'Bio.KDTree', # disabled by default to avoid C++ compilation errors
---
>   'Bio.KDTree', # disabled by default to avoid C++ compilation errors
444c444
<           include_dirs=["Bio/Cluster"]
---
```

```
>           include_dirs=["Bio/Cluster","/usr/local/include/python2.7"]
446,451c446,452
< #   CplusplusExtension('Bio.KDTree._CKDTree', # Disabled by default to avoid
< #           ["Bio/KDTree/KDTree.cpp",       # C++ compilation errors
< #            "Bio/KDTree/KDTree.swig.cpp"],
< #           libraries=["stdc++"],
< #           language="c++"
< #           ),
---
>   CplusplusExtension('Bio.KDTree._CKDTree', # Disabled by default to avoid
>           ["Bio/KDTree/KDTree.cpp",       # C++ compilation errors
>            "Bio/KDTree/KDTree.swig.cpp"],
>           libraries=["stdc++"],
>           language="c++",
>           include_dirs=["Bio/Cluster","/usr/local/include/python2.7"]
>           ),
```

- run following steps:

```
python setup.py build
sudo python setup.py install
```

3. Downloading and running RNAmap2D

- download rnamap2d_linux_py2.7.tar.gz file from ftp://ftp.genesilico.pl/pub/software/rnamap2d/
- change to the directory with the file
- run following steps:

```
tar -zvxf rnamap2d_linux_py2.7.tar.gz
cd rnamap2d_py2.7
python RNAMAP2D.py
```

A1.2 Installation instructions under older Ubuntu/Kubuntu or any other APT-based Linuxes

On other Linux systems, all packages used by the program that are not explicitly programmed by the authors of RNAmap2D, must be installed on the system, BEFORE the installation of RNAmap2D.

Please note that RNAmap2D Linux build supports both python2.6 and python2.7. The user should download the appropriate version from RNAmap2D distribution website, depending on which python version will be used to install and run the program.

Please note that in case of either of python2.6 / python2.7 environment, the user should use the proper interpreter to run the program. As an example, if python2.6 environment is set up to run RNAmap2D, the user should ensure that the python command runs python2.6 interpreter, or run the program using python2.6 command explicitly.

Important Note: The program installation procedures were tested according to the instructions to be found in Appendix A. That is, substituting library versions or switching libraries between python2.6 and python2.7 environment might cause the program not to work. This is why we highly recommend following instructions, including installation of strict library versions.

Table I a; Required packages and versions for the Ubuntu/Kubuntu/APT-based version of RNAmap2D.

name	Ubuntu package name	Python2.6 version	Python2.7 version
Python 2.6 / Python 2.7	python2.6 / python2.7	2.6.4	2.7.3
Python2.6dev / Python2.7dev	python2.6-dev / python2.7-dev	2.6.4	2.7.3
g++	g++	1.4.4	4.6.3
Python Imaging Library (PIL)	python-imaging	1.1.6	1.1.7
Numeric	see: Table I b **IMPORTANT:** after installing, the Lib folder should be copied into python's dist-packages dir as Numeric	24.2	24.2
Numpy	python-numpy	1.3	1.6.1
mxTextTools	python-egenix-mxtexttools	3.1.2	3.2.1
wxPython	python-wxgtk2.8	2.8.10	2.8.12
PyExcelerator	python-excelerator	0.6.3	0.6.4
BioPython (including KDTree sub-package) **installed from source**	see: Table I b **IMPORTANT:** KDTree should be uncommented in setup.py If you have a newer BioPython version installed already, you can copy Bio/ and Martel/ from the build/lib.linux.. directory to the RNAmap2D directory.	1.42	1.4.2
PyCogent	python-cogent	1.4	1.5.1

A1.3 Installation instructions under any other Linux/BSD system

The table I b lists both library versions and URL resources for each library required for installing RNAmap2D. The version information can be combined with table Ia if the user wants to set up python2.7 environment.

Table I b; Required package list and versions for any other linux version of RNAmap2D.

Name	Python2.6 version	download from
Python 2.6	2.6.4	http://www.python.org
Python Imaging Library	1.1.6	http://www.pythonware.com/products/pil/
Numeric	24.2	http://sourceforge.net/projects/numpy/files/Old **IMPORTANT:** after installing, the Lib folder should be copied into python's dist-packages dir as Numeric
Numpy	1.3	http://sourceforge.net/projects/numpy/files/
mxTextTools	3.1.1	http://www.egenix.com/products/python/mxBase/
wxPython	2.8.10	http://sourceforge.net/projects/wxpython/files/wxPython/
PyExcelerator	0.6.4	http://sourceforge.net/projects/pyexcelerator.
BioPython (including	1.42	http://biopython.org/wiki/Download

KDTree sub-package)		**IMPORTANT**: uncomment all lines mentioning 'KDTree' in setup.py before running 'python setup.py build'. If you have a newer BioPython version installed already, you can copy Bio/ and Martel/ from the build/lib.linux.. directory to the RNAmap2D directory.
PyCogent	1.4	http://sourceforge.net/projects/pycogent

g++ and **python2.6-dev** packages are either present or easy to download in any other Linux systems. Please refer to the respective manuals.

Please note, that other library versions may cause RNAmap2D not work.

A.2. MacOSX

RNAmap2D 1.7.x distribution for OSX systems is stand-alone, hence does not requires pre-installation works. In order to install and run the program, it is enough to download the OSX archive, unpack it (this does not require any additional tool) and run the executive file.

A.3. Windows

RNAmap2D distribution for Windows systems is basically stand-alone, hence does not requires pre-installation works. In order to install and run the program, it is enough to download the Windows archive, unpack it (this does not require any additional tool on Windows Vista and Windows 7) and run the executive file.

A.4. RnaView

To enable identification and classification of the types of base pairs that are formed in nucleic acid structures, the user may use third-party program **rnaview**. Rnaview is NOT included in the RNAmap2D package, however the user is encouraged to download and install it and make it visible at the defined path visible for the operating system.

To make RNAView available to RNAmap2D, please install:

rnaview (available at http://ndbserver.rutgers.edu/services/download/)

after the installation create a link to the RNAView binary in /usr/bin by typing: sudo ln -s /?/?/?/RNAVIEW/bin/rnaview /usr/bin

RnaView works on Linux and MacOS systems.

If the user does not install RnaView, RNAmap2D will use its own procedure instead. However it is a prototype module that does not assign cis\trans base pairs reliably – therefore, all of the contacts are displayed as cis.

A.5. Types of distribution

Table III; Type of RNAmap2D distribution

feature	Windows	Linux	OSX
supported	+	+	+
distribution	standalone	prerequisites	standalone
RNAVIEW plugin	-	+	+

Appendix B – Contact map text file formats

B1. File formats

B1.1 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, which equals to the length of the protein. In order to describe the sequence information, first 11 characters corresponding to the "species" name are used by RNAmap2D.

position	value
1	chain id
2 – 5	absolute residue id
6 – 8	residue name
9	space (blank)
10	prot.sec.struct: H(helix), E(strand), -(other), white(missing), blank
11	space (blank)
12 – ...	map records: formatted numbers, delimited by space (see: A3)

B1.2 CASP

This format is used in the **CASP** protein structure prediction experiment since the Contact Map Prediction (B) 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 sequence (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), 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)

B1.3 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 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	nucleotide1 single-letter code
24 – 28	space (blank)
29	nucleotide2 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)

B1.4 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.

CLANS has numerous input/output formats, and the one used by RNAmap2D 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
• **>seq_names**	this section is constructed as in the Phylip case (see: A 1.1)
• **</seqs>**	fixed: sequence section end
• **<mtx>**	fixed: matrix section start
• **matrix**	contact matrix formatted as in the Phylip case (see: A 1.1)
• **</mtx>**	fixed: matrix section end

B1.5 CSV

This is a simple format following a general **comma-separated values** standard.

It contains only the contact map section, which is simply the matrix (see e.g.: A. 1.1, A 1.4), with its values separated by the coma character. This format does not include Summary and statistics information. See also: B.2

B1.6 MS Excel

The 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, iWork Numbers, Google Docs 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.

B.2 Summary and statistics

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

- number of models
- distance metric
- sequence separation (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

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	place against map	placement in file
Phylip	-	below matrix	at the end
Clans	**#** (comment)	above matrix	after length declaration
CASP/EVA	**REMARK** tag	above matrix	after file indicator tag
Excel	sheet name	separate worksheets	'Summary' and 'Statistics'
CSV	N/A	N/A	no summary/stats in file

B3. Values

B3.1 ordinary 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), by a coma character (CSV) or placed in different cells (EXCEL).

B3.2 color saving and retrieving

There are two jobs in RNAmap2D (**single model contact map** and **complex contact map**) where the user can use a special panel to set color to specific contacts. When saving map to a text file, it is later possible to read out the original colors using **contact map from file** job. To achieve this, special color coding was employed. Those values bear no significant meaning.

B3.3 semantics

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 complex contact map
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
colored contact map	color codes	color codes no semantics	contact map colored and saved
sec. structure contact map	0.000, 0.500 and 1.000	1.000 = typical pairing 0.500 = pseudoknot	contact map (encoding) secondary structure