RANSAC: Outlier Detection and Angular Constraining for Subtomogram Alignment in cryoET

Jose-Jesus Fernandez [1], Sam Li [2]

[1] CINN-CSIC, ISPA. Oviedo. Spain.
[2] Dept. Biochemistry and Biophysics. UCSF. USA.

November 2024

Contact: jjfernandez.software@gmail.com

Web site

RANSAC is a software program that analyzes the alignment parameters of subtomograms extracted from continuous filament-like structures (e.g. triplet microtubules in centrioles, doublet microtubules in axonemes, true filaments), detects potentially misaligned subtomograms and estimates their correct alignment from the neighbouring subtomograms by imposing an angular consistency constraint. To this end, the program uses the RANSAC method (RANdom SAmple Consensus) to model the alignment of the filament-related subtomograms and determine the subtomograms that are not consistent with the model (i.e. outliers). The alignment parameters of the outliers are then corrected based on the model calculated from the inliers. The program RANSAC accepts input alignment parameters in Spider/Xmipp format obtained with Xmipp MLTOMO and produces output in the same format. Future versions of the program will accept other formats on demand (e.g. Relion Star), just ask for it at the contact address.

Table of Contents

1 Installation

- Uncompress the file `RANSAC.zip`.
- You will find these directories:
  - `bin`, which contains the executable binaries:
    * `ransac`
      64-bit executable program for Linux. It also works nicely on Windows, using the Windows Subsystem for Linux (WSL) (see below). Minimum processor requirements: a 64-bit Intel or AMD processor.
    * `ransac.osx`
      64-bit executable program for Macs based on Intel or Apple Silicon processors. It was built under OSX 10.11 (El Capitan) on an Intel-based machine. Macs with Apple Silicon require Rosetta 2 (see below).
  - `doc`, with this documentation.
  - `tutorial`, with auxiliary material to run the examples.
- Set up your `PATH` environment variable to have direct access to the executables

RANSAC on Macs with Apple Silicon

RANSAC for Macs was natively compiled on an Intel-based machine. It works smoothly on Macs equipped with Apple Silicon processors provided that Rosetta 2 is installed. If your system does not have Rosetta 2 installed yet, just execute the following command on a terminal to install it:

```
/usr/sbin/softwareupdate --install-rosetta --agree-to-license
```

More information about Rosetta 2 can be obtained in the following links:

https://osxdaily.com/2020/12/04/how-install-rosetta-2-apple-silicon-mac/

https://support.apple.com/en-us/HT211861

RANSAC on Windows (Windows Subsystem for Linux, WSL)

RANSAC runs on command line terminals. The Linux version works nicely on Windows using the Windows Subsystem for Linux (WSL). The reader can find more information about WSL in the following links:

https://learn.microsoft.com/en-us/windows/wsl/install

Installation and execution of RANSAC on WSL is carried out exactly as described in the other sections of this documentation for Linux or OSX.

2 Description

The program RANSAC analyzes the alignment parameters of subtomograms extracted from continuous filament-like structures (e.g. triplet microtubules in centrioles, doublet microtubules in axonemes), detects potentially misaligned subtomograms and estimates their correct alignment from the neighbouring subtomograms by imposing an angular consistency constraint (Fig. 1). Please note that the subtomograms extracted from a filament-like structure are also known as **segments** in this document and in the program.

For the set of subtomograms (segments) from a filament, a linear model of the alignment parameters is fitted following a modified version of the RANSAC method (RANdom SAmple Consensus) [1]. Specifically, there is an exhaustive evaluation of pairs of subtomograms and, for each pair, linear models for the Euler angles (ϕ, θ, ψ) are fitted. Then, the alignment parameters of the whole set of subtomograms in the filament are compared to the models (ϕ, θ, ψ) and the number of outliers is computed (that is, subtomograms whose alignment parameters -ϕ, θ, ψ- are farther than a threshold and hence are inconsistent with the model, see red arrows in Fig. 1). The best models (ϕ, θ, ψ), i.e. with the minimum number of outliers, are finally selected. The outliers are recovered by estimating their alignment parameters (Euler angles and shifts) through regression from the final inliers (green arrows in Fig. 1). If the number of outliers is exceptionally high, the whole set of subtomograms of the filament-like structure are discarded.

Note that the Euler angles do not uniquely specify an orientation (that is, two different sets of Euler angles may specify the same orientation, for instance 90,0,0 and 0,0,90). Internally, the program deals with these situations.

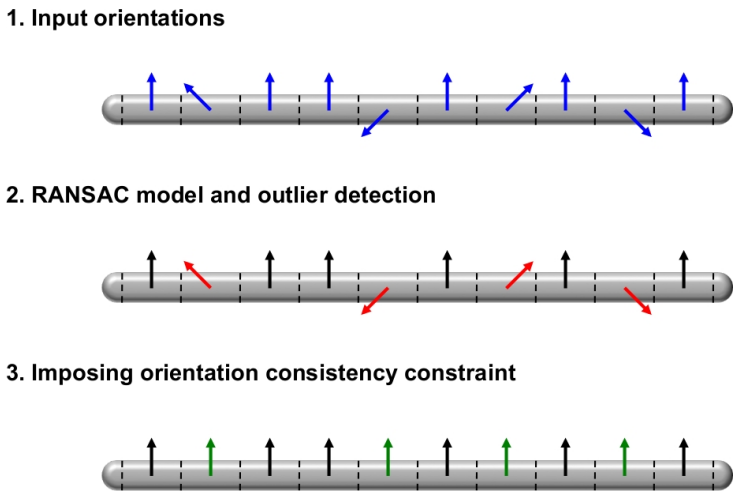

Figure 1: Procedure in RANSAC program. (1) The orientations (blue arrows) for all segments/subtomograms associated to filament-like structures are read from input. The limits of the segments/subtomograms are denoted by the dashed lines. (2) The program RANSAC then models the orientation (actually, three models, for ϕ, θ, ψ) along the filament using the RANSAC method (RANdom SAmple Consensus) [1, 2] applied exhaustively for all pair of subtomograms. The optimal orientation models (ϕ, θ, ψ), with minimum number of outliers (red arrows) and maximum of inliers (black arrows), are selected. (3) Finally, the inliers keep their original orientations (black arrows) whereas the outliers are recovered by estimating their alignment parameters through regression from the final models (green arrows).

If you use RANSAC in your works, please cite any of the following articles where this program has been used:

S Li, JJ Fernandez, MD Ruehle, RA Howard-Till, A Fabritius, CG Pearson, DA Agard, ME Winey.
The structure of cilium inner junctions revealed by electron cryo-tomography
BioRxiv 2024.09.09.612100, 2024
https://doi.org/10.1101/2024.09.09.612100

S Li, JJ Fernandez, AS Fabritius, DA Agard, M Winey.
Electron cryo-tomography structure of axonemal doublet microtubule from Tetrahymena thermophila
Life Science Alliance 5(3):e202101225, 2022
https://doi.org/10.26508/lsa.202101225

S Li, JJ Fernandez, WF Marshall, DA Agard.
Electron cryo-tomography provides insight into procentriole architecture and assembly mechanism
eLife 8:e43434, 2019
https://doi.org/10.7554/elife.43434

3 Usage

This section presents a detailed description of the options and parameters of RANSAC. The program works with a command line user interface that adheres to typical Unix-style practices.

Usage:

```
ransac [options] input output
```

input: The input file is expected to be a DOC file from Xmipp MLTOMO containing the alignment parameters for the whole set of subtomograms from all filament-like structures in a project.

output: The output file is a DOC MLTOMO file with the alignment parameters of the subtomograms that are kept: inliers as well as outliers that were recovered. The alignment parameters of the outliers are overwritten with those estimated by regression from the inliers.

RANSAC also generates two output SEL files (Xmipp format) with the names of:

- the discarded subvolumes: output_discarded.sel
- the kept subvolumes: output.sel

The latter SEL file allows the user to quickly run a MLTOMO command with only the subtomograms that have consistent alignment parameters (i.e. inliers and recovered outliers).

Input/Output file format

The DOC format used for the input and output MLTOMO files is Spider/Xmipp and has:

- A header line at the beginning of the file (starting with ' ;').
- For each segment/subtomogram, there are two lines:
  - a line starting with ' ;' and containing the file name of the subtomogram.
  - a line with 12 columns, with the first 8 columns being:

    key registers_per_line ϕ θ ψ X_offset Y_offset Z_offset

    with ϕ,θ,ψ being the Euler angles and X_offset,Y_offset,Z_offset being the shifts determined in the alignment.

    Actually, the RANSAC program only works with the 6 columns ϕ,θ,ψ,X_offset, Y_offset,Z_offset from the input MLTOMO file. The final alignment parameters after applying the RANSAC procedure are written to the output file in the proper 6 columns and the remaining columns to complete the total of 12 are just reproduced from the input file.

Options to control the RANSAC procedure

-a <angle>
Angular threshold to consider a subtomogram as an outlier. This parameter sets the angular difference between the original input Euler angles assigned to the subtomogram and the models fitted (ϕ, θ, ψ). If any of the differences in ϕ, θ, ψ is higher than ($\geq$) the threshold, the subtomogram will be considered as an outlier. By default, a value of 10.0 degrees is taken.

-c <fraction>
Minimum fraction of segments in a filament to consider internal consistency. If the fraction of inliers for the best fitted model is lower than this value, the whole filament is considered inconsistent and all segments are discarded. By default, a value of 0.5 is taken, that is, at least 50% of the segments in a filament must be consistent with the best fitted model (i.e. they must be inliers).

-C <NSegments>
Minimum number of segments to check outliers. If the number of segments in a filament is lower than this value, RANSAC is not applied and the original alignment parameters are preserved for all segments in this filament. By default, a value of 3 segments is taken.

-d <angle>
Maximum angular deviation -in average- of contiguous segments. By default, a value of 2.5 degrees is taken. This parameter is used during the model fitting in RANSAC to impose smooth differences (i.e. small slope in the fitting) in alignment parameters through the length of the filament. This value is taken on average with the number of segments, so if this value is 2.5 degrees and the number of segments is 10, the maximum difference between the first and last segment should be less than 25 degrees. If the value of this parameter is too small, it might be too restrictive and the number of outliers will increase. This parameter is case-dependent, for instance in a filament where the segments are significantly overlapped, the value may be small. However, if there is no overlapping between the segments, the value should perhaps be larger to acommodate larger angular differences between consecutive segments.

-l <log_file>
Option to produce an output log file with a detailed description of the procedure for all segments of all filaments: the final alignment parameters, the original ones, whether they were considered outliers and their new alignment parameters obtained through regression or whether they were discarded.

-s <NSegments>
Number of segments into which the filament-like structures were divided in a fixed manner. Thus, using this option assumes that all filaments were divided into the same number of segments.

-S <seg_file>
Input file with information about the number of segments when they are non-fixed, that is, when each filament may be divided into a number of segments that is different from other filaments. This file allows the user to uniquely specify the indices of the tomogram,

filament and segment associated to each alignment parameter in the input DOC file (see above).

The format of this file is Spider and it is assumed to have:

- A header line at the beginning of the file (starting with ' ;').
- a line per segment/subtomogram with at least the following 5 columns (there can be more columns, but they will be ignored):

  key registers_per_line tomogram_index filament_index segment_index

  The first two columns (key and registers_per_line) are directly ignored. The order of the segments/subtomograms in this file are assumed to be exactly the same as in the input DOC file with the alignment parameters of the whole set of subtomograms (see above). So, the alignment parameters of the i-th subtomogram in the input DOC file correspond to the indices (tomogram, filament, segment) presented in the i-th order in this file.

-t <angle>

Threshold on θ to treat it as 'near-zero'. The procedure in RANSAC is slightly different for those alignment parameters with θ close to zero so as to properly deal with some of the non-uniquely defined orientations by the Euler angles. In particular, the decision whether a segment is an outlier is made depending on θ and $\phi + \psi$. Those θ angles lower than this threshold are treated this way. By default, a value of 30.0 degrees is taken.

-v

Flag to activate verbose execution in order to obtain some statistics of the input subtomograms and filament-like structures. By default the program works silently.

-w

Flag to activate writing the angles of the alignment parameters not strictly wrapped in a 360-range. This is useful to see the trends that were internally used to fit the models and identify the outliers.

4 Examples

In the directory `tutorial` there are several examples of MLTOMO alignment files to practice with RANSAC.

1. First, there is an MLTOMO alignment file from 1000 subtomograms of Basal Body extracted from triplet microtubules (TMT) divided into 10 segments: `mltomo_bb1000.doc`. So the RANSAC command must use the option '`-s 10`' to properly interpret the segments and proceed with the model fitting. The RANSAC command could be:

```
ransac -v -l mltomo_bb1000_ransac.log -s 10 mltomo_bb1000.doc
mltomo_bb1000_ransac.doc
```

The output MLTOMO file (`mltomo_bb1000_ransac.doc`) along with the (silently created) output SEL file (`mltomo_bb1000_ransac.sel`) can be directly used in a new round of MLTOMO to further refine the alignment. In addition, through the use of '`-l`', an output log file is generated (`mltomo_bb1000_ransac.log`) with a detailed description of the RANSAC procedure for each subtomogram, indicating the final alignment parameters (either the original parameters if it is an inlier, or the estimated ones through regression if it is an outlier) or it has been definitely discarded.

2. Second, there is an MLTOMO alignment file from 800 subtomograms of Pro-Basal Body extracted from triplet/doublet microtubules (TMT/DMT) divided into a non-fixed number of segments: `mltomo_pbb0800.doc`. In this case we need to have an additional file describing the division of the filament-like structures into segments: `pbb0800_indices.spi`, where the last three columns represent the indices of the tomogram, filament (specifically TMT/DMT in this case) and segment. The RANSAC command then needs to input this file with the option '`-S pbb0800_indices.spi`':

```
ransac -v -d 6 -l mltomo_pbb0800_ransac.log -S pbb0800_indices.spi
mltomo_pbb0800.doc mltomo_pbb0800_ransac.doc
```

In this case, the default threshold (2.5 degrees) on the maximum angular deviation between contiguous segments seems to be too restrictive, raising outliers that may be actually false. While this is not usually harmful, it is convenient that these subtomograms are considered as inliers so that their original alignment parameters are preserved (otherwise, the parameters would be obtained through regression). So we used the option `-d 6` to increase the threshold to reduce the number of outliers and preserve their original alignment parameters.

Also, an output log file with a detailed report of the RANSAC procedure is generated using '`-l mltomo_pbb0800_ransac.log`'.

References

[1] M.A. Fischler, R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. doi: https://doi.org/10.1145/358669.358692

[2] https://en.wikipedia.org/wiki/Random_sample_consensus

Acknowledgements

The development of RANSAC has been supported by HHMI, NIH, Fundacion Ramon Areces and Spanish AEI.