SHAPES 1.0/1.2/1.3 Documentation Revised 9-June-2019

Overview

The SHAPES program calculates molecular envelopes for protein samples from x-ray solution scattering data. The envelope is determined by matching the experimentally determined P(r) (the particle or pair distance distribution function), as calculated by the program GNOM from the ATSAS suite, to the P(r) obtained from a constellation of volume-filling beads. The final arrangement of beads is subject to the condition that the beads fill out the expected protein volume and are distributed across that volume in a relatively uniform fashion.

In tests performed with synthetic data, with intensities calculated from a wide range of protein models, SHAPES was able to determine molecular envelopes consistent with the input models. Reconstruction calculations with experimentally determined P(r), previously used with other shape reconstruction software, also gave correct reconstructions when used as inputs for SHAPES. SHAPES runs relatively quickly and most reconstruction runs should complete within a few minutes. The SHAPES program typically gives very similar reconstructions in replicated runs, obviating the need to average multiple reconstructions to obtain the final molecular envelope.

The SHAPES distribution includes the SHAPES program (*shapes_v#.py*), a GNU GPLv3 license (*COPYING.txt*), and an example input file (*shapes_ip.txt*). The most current version of this documentation and a summary of the terms and conditions for use of the program are available from internet site **http://saxs2shapes.com**.

How to run SHAPES

The SHAPES program is currently distributed as single file, *shapes_v#.py*, in Python code. The program was developed and tested on the Windows operating system using python 2.6.5 but should execute correctly on any computer system on which a python 2 interpreter is installed. SHAPES versions 1.2/1.3 have been modified and tested using python 3.7 and should be compatible with both python 2 and python 3. SHAPES may be launched via a double-click on the program icon (from the Windows operating system) or from the command line. For example,

'your-path-to-python-interpreter' 'your path to the *shapes_v#.py* program'

The operation of SHAPES is controlled by a keyword file, *shapes_ip.txt*, that should be present in the folder in which the SHAPES program is executed. The SHAPES input parser expects to read lines in this file that contain the keyword as the first item and the associated parameter as the second item.

In addition to the terminal output, summary output from each run is also captured in an output file, *shapes_summary.log*.

Keyword Input

Only two parameters, specified by the **num_amino_acids** and **input_pr** keywords are required to run SHAPES. All other parameters will be set to default values if they not provided in the *shapes_ip.txt* file.

All keywords are followed by a single parameter.

Possible keywords are:

id (Added in version 1.2/1.3). An option to add a prefix to all of output pseudo-pdb and data file names. This option may be useful for distinguishing sets of output files resulting from different runs carried out within the same folder.

num_amino_acids The total number of amino acids in the structure, regardless of symmetry or subsequent adjustment to the number of beads employed in calculations. As a special case, this parameter must be set to zero if the starting positions are to be read from a file of Cα positions specified by the **pdb_start** parameter.

input_pr The P(r) file, as output by GNOM program (*gnom.out*) from the ATSAS package

num_solns The number of reconstruction runs. Typically, this parameter could be set to 4-10 for production runs that allow analysis of the variability in structure solutions across multiple reconstruction trials.

num_aa_scale A scale (a divider) on the number of amino acids (**num_amino_acids**) that may be used to alter the number of volume-filling beads used in calculations. Typically, this parameter is set to 1.0 and the number of particles used to fit the P(r) is equal to the expected number of amino acids. Setting this parameter to a value less than unity will produce more (smaller) beads than amino acids. It is recommended that reconstruction calculations include at least ~300 beads so, for example, for a small protein containing ~200 amino acids this parameter might be set to ~0.67. Setting this parameter to a value greater than unity causes the program to employ fewer (but larger) beads than the number of amino acids. By using a reduced number of beads the reconstruction calculation will be greatly accelerated and this option can be very useful for allowing SHAPES to function efficiently on larger proteins and protein oligomers (> 1000 amino acids in the symmetric unit).

symm The point rotational symmetry for protein assemblies (i.e 1 for no symmetry, 2 for a homo-dimer etc). This option is recommended when rotational symmetry in the protein complex is anticipated. The z-axis is the rotation axis in the resulting pseudo-PDB output files.

bias_z A factor to bias the starting configuration of beads along the Z-axis (default is usually 0.0 for no bias). Positive values increase the allowed range of the initial distribution of beads along Z and reduces the distribution in XY to give a more rod-like starting point. Setting negative values has the opposite effect and produces a more disk-like starting point. This option is mainly intended for use in conjunction with the specification of point rotational symmetry (> 2) to explore starting points that are set to be rod-like or disk-like (suggested values, 0.2 or -0.2) and avoid reconstructions where the symmetry axis has become trapped along the 'wrong' dimension.

inflate_vol A scale factor (default 1.0) for adjusting the starting and final volumes relative to the expected partial specific volume. From SHAPES v1.1 onwards the bead-separation distance is also proportionately inflated to match the inflated final volume. For low resolution envelopes recovered from SAXS reconstructions the partial specific volume tends to cover a tight, minimal core volume since proteins contain many small cavities that become filled in the reconstruction and contribute to the

assessment of the protein volume. For loosely organized proteins or proteins that contain a large proportion of beta-sheet structure it may be found preferable to generate solutions that cover a somewhat larger volume. Excessive variability in structure solutions may also signal that the target volume is too small. Moderate inflation factors (1.2-1.4) may be applied for this purpose.

pdb_start An option to initiate the reconstruction process using the CA atoms read from an input PDB file. The PDB file name is the input parameter. Options to apply symmetry (**symm**) and to adjust the number of beads (**num_aa_scale**) are not operable when this option is used. The parameter **num_beads_start** must also be set to 0 to invoke this option. This option is intended to explore small variations in protein shape around the form of the input model. From version 1.2 onwards, Input bead positions are not subject to a preliminary energy minimization when this option is used.

glue An option (default value 0.0) introduced in SHAPES version 1.3 to use for 'difficult' reconstructions (extremely elongated or flattened protein shapes) where the initial reconstruction trials may have given broken volumes or disparate results. This option stabilizes the reconstruction procedure by adding an energy penalty that prevents cavities opening up in the protein volume and usually results in reconstructions that have more contiguous volumes and that more similar to each other. If the differences between a set of reconstructed volumes (files *psv_shape_#.pdb*) are large (NSD >~ 0.70 over non-outlier reconstructions as assessed by the damsel function from the ATSAS suite) it may be worth trying a glue weight of 5.0. For extremely elongated proteins where the reconstruction runs resulted in broken volumes and glue weight of 10.0 might be tried.

Outputs and recommended use

SHAPES is intended to output reliable and reproducible reconstructions of protein molecular envelopes without further processing or manipulation. It is recommended that a small number of shape determination runs (5-10, set via the **num_solns** keyword) are made to assess the variability in the resulting reconstructions. The variation in the resulting molecular volumes should be low except for occasional outlier solutions. The most representative reconstruction might be selected from this set for display and analysis.

SHAPES outputs the pseudo-PDB files *beads_#.pdb* and *psv_shape_#.pdb*, where # denotes the run number up to the value specified by the **num_solns** keyword. The files *beads_#.pdb* report the final constellation of beads used to fit the input P(r). The predicted molecular envelope from each run, typically slightly larger than the expected partial specific volume, is contained in the file *psv_shape_#.pdb* and represented by an array of dummy CA atoms set on a 5 Å grid.

For most purposes the most representative member of the *psv_shape_#.pdb* files will be the useful output to compare to atomic models.

Evaluation of solutions

Numerical scores for evaluating solutions are the difference between the measured P(r) and the P(r) modeled by beads and a VDW energy (E) for interactions between beads. Typical values for successful runs are 0.05 and -1.4 respectively. Reconstructions from data on small proteins may sometimes result in somewhat higher P(r) scores. These scores are reported in the headers of the output files (*psv_shape_1.pdb* etc) and in a summary log file, *shapes_summary.log*.

From SHAPES version 1.2 onwards additional diagnostics are provided by output files *pr_calc_#.dat* and *intensity_#.dat.* These files contain the observed and final calculated values for the P(r) target and intensity data respectively. The calculated P(r) is typically somewhat jagged due to the limited number of particle pairs and the use of points to represent bead positions but should follow the shape and range of the input P(r). The calculated intensities should agree well with the observed data at low-q but may deviate somewhat towards the end of the data range.

These statistics only provide a sanity check on the reconstruction run. Aberrant (outlier) reconstructions occasionally occur and confidence and are best identified by comparing solutions from a small number of runs. A hallmark of successful reconstruction calculations is that solutions from multiple runs are consistent. If the NSD score, as run via the 'damsel' command, as incorporated in the SUPCOMB program from the ATSAS suite, is used to compare structure solutions, the mean value for pair-wise comparisons across all files of type *psv_shape_#.pdb* should be significantly smaller than unity and typically lies in the range 0.5-0.7 for reliable reconstructions. The **glue** keyword is available to try for more difficult cases if the results of initial trials are unsatisfactory.