

Mesa ShapeBase Toolkit

This document provides an overview of the installation and use of the Mesa Analytics shape tools.

Installation

Unpack the distribution tarball.

```
$ tar xzf mesaac_Linux-2.6.22.5-31-default-x86_64-with-glibc2.3.tgz
```

Install the Mesa eval license file.

Note: The license file is shipped as 'mesaac_license', but it must be installed as '.mesaac_license' (with a leading dot).

```
$ cd mesaac_Linux-2.6.22.5-31-default-x86_64-with-glibc2.3
$ cp mesaac_license ~/.mesaac_license
```

Programs

Shape Fingerprinter

shape_fingerprinter generates shape fingerprints for the 3D conformers in *sd_file*. The shape fingerprints are written to standard output.

Usage

```
shape_fingerprinter [options] sd_file hamms_sphere_file atom_scale
```

sd_file

a file of conformers in SD format, with 3D coordinates

hamms_sphere_file

a file containing 3D Hammersley sphere points, one point per line with space-separated coordinates, for principal axes generation via SVD and fingerprint generation

atom_scale

the amount, in the range [1.0 .. 2.0], by which to increase atom radii for alignment

shape_fingerprinter options:

-i | --id

if specified, include the name of each SD conformer after each fingerprint,
separated by a space

-c | --compress

write fingerprints as gzipped, base64-encoded binary-ascii strings

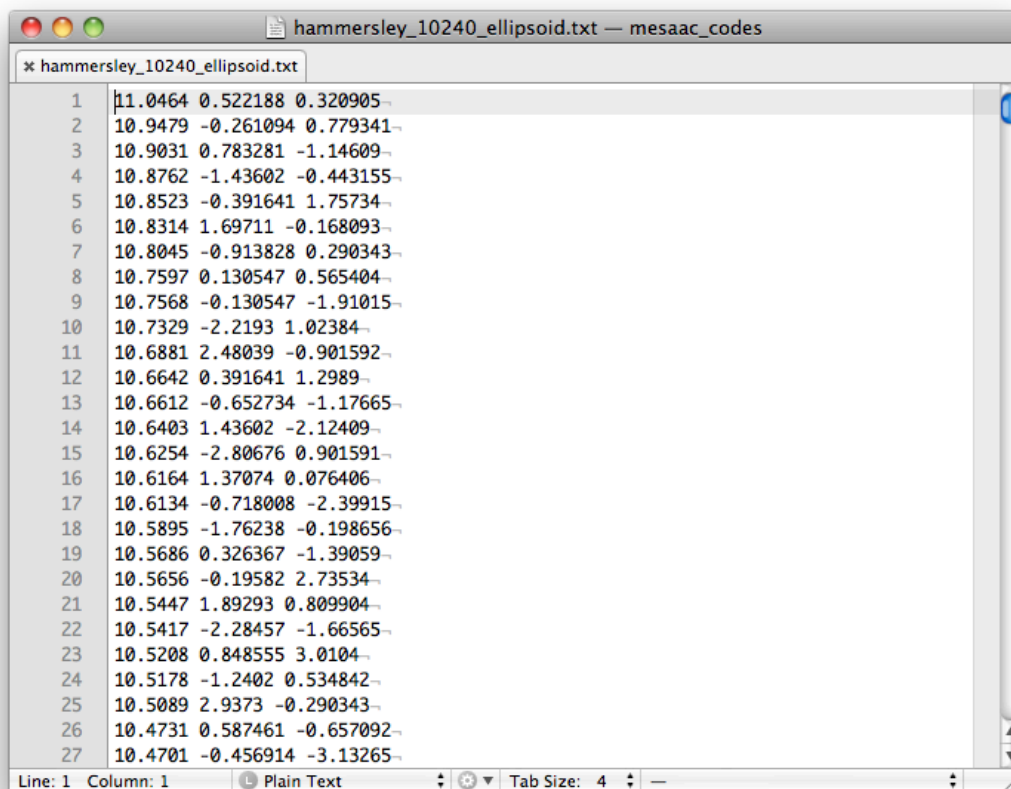
-e | --ellipsoid *ELLIPSOID_FILE*

use points from *ELLIPSOID_FILE*, a file containing 3D Hammersley ellipsoid
points, one point per line with space-separated coords, for fingerprint generation

-h | --help

print this help message and exit

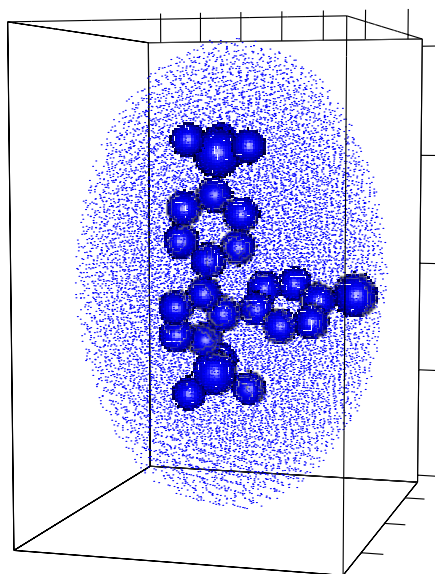
The Hammersley ellipsoid and sphere files consist of plaintext lines, each line containing whitespace-separated floating point coordinates of a 3-space point.



```
hammersley_10240_ellipsoid.txt
1 11.0464 0.522188 0.320905
2 10.9479 -0.261094 0.779341
3 10.9031 0.783281 -1.14609
4 10.8762 -1.43602 -0.443155
5 10.8523 -0.391641 1.75734
6 10.8314 1.69711 -0.168093
7 10.8045 -0.913828 0.290343
8 10.7597 0.130547 0.565404
9 10.7568 -0.130547 -1.91015
10 10.7329 -2.2193 1.02384
11 10.6881 2.48039 -0.901592
12 10.6642 0.391641 1.2989
13 10.6612 -0.652734 -1.17665
14 10.6403 1.43602 -2.12409
15 10.6254 -2.80676 0.901591
16 10.6164 1.37074 0.076406
17 10.6134 -0.718008 -2.39915
18 10.5895 -1.76238 -0.198656
19 10.5686 0.326367 -1.39059
20 10.5656 -0.19582 2.73534
21 10.5447 1.89293 0.809904
22 10.5417 -2.28457 -1.66565
23 10.5208 0.848555 3.0104
24 10.5178 -1.2402 0.534842
25 10.5089 2.9373 -0.290343
26 10.4731 0.587461 -0.657092
27 10.4701 -0.456914 -3.13265
```

The Hammersley ellipsoid file contains the points from which fingerprints are generated. If a given ellipsoid point lies within the volume of a conformer, then the corresponding fingerprint bit is set.

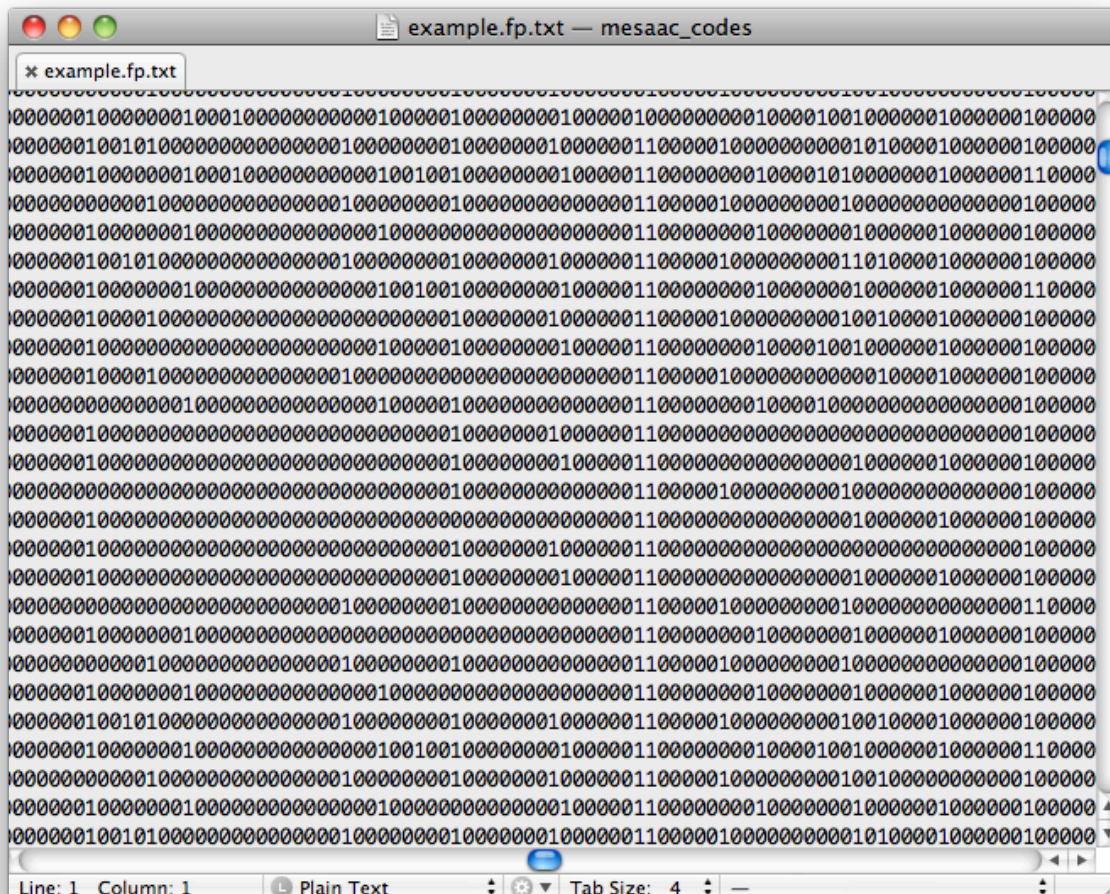
The ellipsoid points define an ellipsoidal volume whose three principal axes lie along the x, y and z coordinate axes, respectively. (Note, the atom radii in this figure are covalent bond radii and not Bondi radii. Also, for effect, the ellipsoid represented here is more elongated along the principal axis than are the ellipsoids used to generate the fingerprints which are more spherical in shape, though still scalene ellipsoids.)



The Hammersley sphere points are used to find the principal axes of the conformers, so they can be aligned to the coordinate axes before generating shape fingerprints.

Depending on the density of points in the two point files, it may be useful to specify an `atom_scale` value greater than 1.0. Doing so causes the atom radii of each conformer to be scaled up by the indicated amount, and may yield denser shape fingerprints.

The output consists of lines of plaintext "bits" (ASCII "0" and "1" characters). Each line contains a single "bit" for each point in the Hammersley spheroid file.

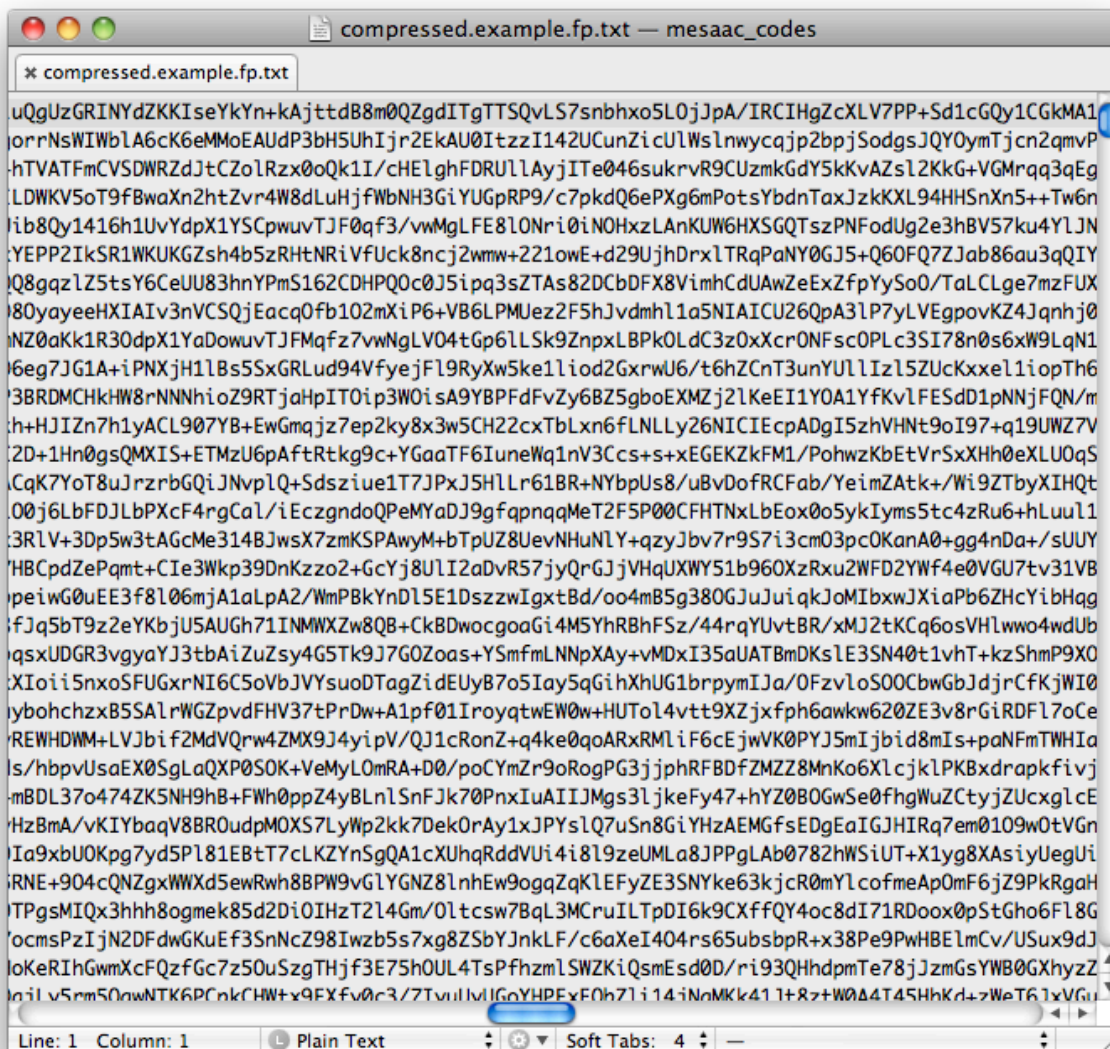


The screenshot shows a text editor window titled "example.fp.txt — mesaac_codes". The editor contains a single line of 1,000,000 characters, each being either a '0' or a '1'. The status bar at the bottom indicates "Line: 1 Column: 1", "Plain Text", and "Tab Size: 4".

Each conformer in the sd input file is represented by four output lines, one for each possible orientation of the conformer with respect to the coordinate axes.

When either `-c` or `--compress` is specified, each output fingerprint is represented as a base64-encoded, gzipped string. When decoded, this string will yield the original "0s and 1s" fingerprint

Use of the `-c` option typically reduces output size by more than an order of magnitude.



When the `-i` (or `--id`) option is specified, an extra space-separated column is appended to each output line, containing the name of the corresponding SD structure. (Note: `shape_fingerprinter` does not check for duplicate or empty names.)

By default `shape_fingerprinter` uses the *hamms_sphere_file* points for both axis alignment and fingerprint generation. The `-e` (or `--ellipsoid`) option lets you specify an *ELLIPSOID_FILE* whose points correspond to the fingerprint bits. See the

hammersley_spheroid program below for more information on generating both spheroid and ellipsoid point sets.

Example

Note: stderr output is indicated in italics.

```
$ bin/shape_fingerprinter conformers.sd \  
  data/hammersley/hamm_spheroid_20k_11rad.txt 1.0 >fingerprints.txt
```

Running ShapeFingerprinter

Source code Copyright (c) 2009

Mesa Analytics & Computing, Inc.

Version number 1.0 Creation Date: April 30, 2009

Expiration Date: Sun Aug 1 01:00:00 2010

```
$ fgrep -c '$$$$' conformers.sd
```

467

```
$ wc -l fingerprints.txt
```

1868

Note: 467 X 4 = 1868: there are 467 conformations in conformers.sd, and each conformation has 4 subfingerprints in fingerprints.txt

Other Programs

Shape Measures

shape_measures takes shape fingerprints as input and outputs various forms of NXN comparison matrices for clustering or MXN comparison matrices for similarity searching.

Usage

*shape_measures fingerprints measure similarity format searching search_number |
alpha | sparse_threshold*

fingerprints

a file of fingerprints as output by shape_fingerprinter

measure

'-T' for Tanimoto, '-V' for Tversky, '-C' for Cosine

Note: Cosine is the Ochiai similarity.

similarity

'-S' for similarity, '-D' for dissimilarity

format

'-M' for Matrix, '-O' for Ordered Pairs, '-S' for Sparse Matrix, '-P' for PVM

Note: '-S' and '-P' options are intended for use with Mesa's Grouping Module input.

searching

'-T' if performing similarity searching, '-F' if not

search_number

a positive integer M, such that $M < N$, to be searched. If *searching* is '-T' then each of the first *search_number* (M) entries in *fingerprints* will be compared to each of the N non-search entries in *fingerprints*; the output will show pairwise distances between each of these M "search" entries and the subsequent N entries. **Note:** *search_number* is always required, though its value is ignored when *searching* is '-F'.

alpha

Tversky alpha parameter, to be specified when measure is '-V' (Tversky). Value should be in the range(0, 2).

Note: $\beta = 2 - \alpha$; if $\alpha = \beta = 1$, '-V' will produce the same results as the '-T' measure

sparse_threshold

sparse distance threshold, in the range of (0, 1). Must be specified when format is '-S' or '-P'.

The various formats are all simple plaintext ASCII.

Example

Note: stderr output is indicated by italics.

```
$ shape_measures fingerprints.txt -T -S -M -F 0
```

Running shape_measures

Source code Copyright (c) 2009

Mesa Analytics & Computing, Inc.

Version number 0.1 Creation Date: May, 2009

Expiration Date: Sun Aug 1 01:00:00 2010

Number of fingerprints is 16

```
1 0.713415 0.72305 0.818854 0.55099 0.91548 0.608263 0.805506 ...
0.713415 1 0.966667 0.653096 0.657437 0.707692 0.708778 0.743723 ...
...
```

Shape Volume

shape_volume takes as input a multiconformer sd file with 3D coordinates and outputs the corresponding volumes, given a CPK van der Waals model with Bondi radii.

Usage

```
shape_volume sd_file hamms_sphere_file sphere_radius atom_scale
```

sd_file

a file of conformers in SD format, with 3D coordinates

hamms_sphere_file

a file containing 3D Hammersley sphere points, one point per line with space-separated coordinates, for principal axes generation via SVD

sphere_radius

radius of Hammersley sphere points (see *hammersley_spheroid* options)

atom_scale

factor by which to increase/decrease atom radii, relative to their van der Waals radii

Example

```
$ shape_volume conformers.sdf data/hammersley/hamm_ellipsoid_20k_11rad.txt
11.0 1.0
270.869
254.808
255.08
...
```

Note: the radius of the 20480 point generated sphere is 11.0 Angstroms.

Align Monte

align_monte reads a multi-conformer sd file with 3D coordinates and aligns all conformations to the first conformation in the file (aka the reference conformer). It writes the aligned conformations to standard output.

Each output conformer includes "<MaxAlignMeasureName>" tags for the specified similarity measure(s), indicating the value of the measure(s) for the best alignment of the conformer with the reference conformer. The measure values for the reference conformer will be "1".

Each output conformer also includes "<BestFlipMeasureName>" tags for the specified measure(s), indicating which axis-flipped orientation of the conformer provided the best match with the reference conformer, for the corresponding measure. Best-flip values should be integers 1-4, inclusive.

By default *align_monte* performs alignment by axis-aligning the set of Hammersley sphere points which lie within the volume enclosed by each conformer's atoms. When the *-a* (or *--atom-centers*) option is specified, *align_monte* uses only the center points of the conformer atoms to perform axis alignment. This option can significantly reduce runtime, but occasionally it reduces the quality of the alignment.

Usage

`align_monte [options] sd_file hamms_sphere_file atom_scale`

sd_file

a file of conformers in SD format, with 3D coordinates

hamms_sphere_file

a file containing 3D Hammersley sphere points, one point per line with space-separated coordinates, for principal axes generation via SVD

atom_scale

the amount, in the range [1.0..2.0], by which to increase atom radii for alignment

align_monte options:

`-al--atom-centers`

if specified, perform alignment using atom centers only; otherwise use contained Hammersley sphere points to perform alignment.

`[-ml--measure B | T | C | V ALPHA [-ml--measure ...]]`

the measures to use in finding the best alignments (B=BUB, T=Tanimoto, C=Cosine, V=Tversky)

Note: If you specify Tversky ('-m V', you must also provide an *ALPHA* value in the range 0..2 inclusive. If no measure option is specified, Tanimoto will be used. For each specified measure type, the SDF conformers will be tagged with corresponding `<MaxAlignMeasureName>` and `<BestFlipMeasureName>` properties

`-sl--sort SORT_FILE`

Create the named *SORT_FILE* containing SDF record indices and corresponding measure values, in descending value order. The last specified measure will be used as the sort value.

`-h | --help`

print this help message and exit

Example

\$ align_monte conformers.sdf hamm_20480_sphere.txt 1.0

Where the reference conformation input is:

1-1

Cerius2 12120216093D 1 1.00000

Structure written by MMmdl.

39 41 0 0 0 0 0 0 0 0999 V2000

27.7051 22.0403 17.0243 C 0 0 0 0 0 0

26.4399 22.0976 16.4318 N 0 0 0 0 0 0

25.5381 21.4424 17.2831 C 0 0 0 0 0 0

....

23 39 1 0 0 0

M END

\$\$\$\$

...and the reference output:

1-1

_Mesaac_06251013083D

Structure written by MMmd1.

39	41	0	0	0	0	0	0	0	0	0999	V2000						
3.0857	2.0564	-0.0566	C	0	0	0	0	0	0	0							
2.0482	1.1209	0.0009	N	0	0	0	0	0	0	0							
0.8376	1.8269	0.0594	C	0	0	0	0	0	0	0							

...

M END

> <MaxAlignTanimoto>

1

\$\$\$\$

A typical aligned structure to the reference will have <MaxAlignMeasureName> values significantly less than 1.0:

1-4

_Mesaac_06251013293D

Structure written by MMmd1.

36	38	0	0	0	0	0	0	0	0	0999	V2000						
2.0617	1.4067	0.0688	N	0	0	0	0	0	0	0							
0.7920	1.9811	-0.0867	C	0	0	0	0	0	0	0							
0.9602	3.3591	-0.1451	C	0	0	0	0	0	0	0							

...

M END

> <MaxAlignTanimoto>

0.732194

\$\$\$\$

Ancillary Programs

The following programs are for those who would like to experiment with the efficiency and accuracy of the quasi-monte carlo aspect of the previous programs; or, they conformation sets of compounds that fall outside the usual volumes or dimensions of the usual drug or drug-like compounds (e.g., fragment libraries, or natural product libraries with large compounds).

Hammersley Spheroid

Given a number of points to be generated within a cube, together with a, b, c parameters specifying the relative extents of the axes of an ellipsoid (with a sphere being the special case of $a=b=c=1$), and a radius parameter in angstroms, `hammersley_spheroid` produces a sequence of quasi-random points lying within the specified ellipsoid volume.

For example, "`hammersley_spheroid 10000 1 1 1 10`" will produce the subset of 10000 points within a square of side length 20, centered about the origin, which are enclosed by a sphere with radius 10. Changing a, b, and c so they are not equal will produce a scalene ellipsoid. Generation of quasi-random points directly within a sphere, without the rejection method, is a known hard problem.

By convention, when producing scalene ellipsoids for input to the shape fingerprint, $a > b > c$, where (a,b,c) correspond to the (x,y,z) axes.

Usage

`hammersley_spheroid sample_size a b c scale`

sample_size

total number of 3D points to generate

a b c

ellipsoid axis scale parameters (continuous values)

scale

extent of largest ellipsoid axis

Example

```
$ hammersley_spheroid 10000 1 1 1 10
9.902 -0.9375 -0.617285
9.812 0.46875 0.288066
9.716 1.17188 -1.68724
9.71 -0.703125 2.01646
9.668 2.10938 0.946502
9.524 0.703125 -1.02881
...
```

Shape Radius

shape_radius takes as input an sd file and prints the maximum extent of any conformation with respect to the mean centered conformations. This program tests the effective size of a database of molecules to determine if the fingerprinting Hammersley spheroids are sufficiently large or small to effectively cover all conformations, while maintaining sufficient density of the sampling points.

Usage

shape_radius *sd_file hamms_sphere_file atom_scale*

sd_file

a file of conformers in SD format, with 3D coordinates

hamms_sphere_file

a file containing 3D Hammersley sphere points, one point per line with space-separated coordinates, for principal axes generation via SVD

atom_scale

the amount, in the range [1.0..2.0], by which to increase atom radii for alignment

Example

Note: stderr output is indicated by italics.

```
$ shape_radius conformers.sdf data/hammersley/hamm_ellipsoid_20k_11rad.txt
1.0
```

Running shape_radius

Source code Copyright (c) 2009

Mesa Analytics & Computing, Inc.

Version number 1.0 Creation Date: April 30, 2009

Expiration Date: Sun Aug 1 01:00:00 2010

```
Max_x Max_y Max_z Max_atom_radius Max_radius_plus_max_atom_radius
8.72715 7.25674 8.51959 1.7 11.7392
```