Piramid

- Version 1.0.0 -

Last update: September 15, 2010

Copyright 2007-2010

Silicos NV

Table of contents

©2007-2010 Silicos NV

1. Introduction

Copyright (C) 2007-2010 by Silicos NV

This file is part of the Open Babel project. For more information, see http://openbabel.sourceforge.net/

This program is free software; you can redistribute it and/or modify it under the terms of the GNU General Public License as published the Free Software Foundation version 2 of the License.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.

Piramid is a program for the alignment of a reference molecule against a database of molecules using the shape of the molecules. It is based on the use of Gaussian volumes as descriptor for molecular shape as it was introduced by Grant and Pickup[1].

Piramid is a command line-driven program that is instructed by means of command line options. The program expects one reference molecule with its three-dimensional coordinates and one database files containing one or more molecules in three dimensions. The results are either the alignment of all database molecules and their respective scores or the N best scoring molecules from the complete database.

[1] Grant J.A., Pickup B.T., 1995. J. Phys. Chem., *A Gaussian Descriptor of Molecular Shape, 99, 3503-3510.*

2. Background

Piramid is based on the computation of the overlap between two molecules when their atomic volumes are represented using a Gaussian function. The rationale is to find the alignment of the reference and database molecule that optimizes their overlap. In the next sections, the methodology is explained in more detail.

2.1. Atom Gaussians

The basis of the shape-based alignment procedure is the modeling of the atomic volume with a Gaussian descriptor. The atom Gaussian is modeled with the parameters *m*, the center or position of the atom, and α, the spread. α is chosen inverse proportional to the squared radius of the atom.

The volume of an atom is in this representation computed as:

$$V_a = \int p\exp(-\alpha \mid m - r \mid^2)dr = p\sqrt{\left(\frac{\pi}{\alpha}\right)^3}$$

The scaling factor *p* is set such that the Gaussian volume is equal to the hard sphere volume of $\frac{4\pi}{3}R^3$.

2.2. Gaussian volume

The total volume of a molecule can be written in terms of atom volumes and higher-order overlaps as:

$$V = \sum_a V_a - \sum_{a,b} V_{ab} + \sum_{a,b,c} V_{abc} - \sum_{a,b,c,d} V_{abcd} + \ldots$$

The interesting part of using Gaussians as representation for atoms is that the overlap between two atom Gaussians (a, α) and (b,β) will give a new Gaussian with the following properties:

- The center is $m = \frac{\alpha a + \beta b}{\alpha + \beta}$;
- The spread is $\gamma = \alpha + \beta$.

The overlap volume itself is given by the following formula:

$$V_{ab} = pp\exp\left(-\frac{\alpha\beta}{\alpha+\beta}\|a-b\|^2\right)\sqrt{\left(\frac{\pi}{\alpha+\beta}\right)^3}$$

Since the overlap of two Gaussians is again a Gaussian it is possible to implement a iterative procedure to compute the higher level overlaps as needed in the volume estimation. The procedure starts with representing all atoms and all atom-atom overlaps as Gaussians. In the next step, extending the atom-atom overlaps with all atoms generates the next level of overlaps. This procedure is repeated up to a given level.

©2007-2010 Silicos NV

The outlined procedure is combinatorial in nature; therefore, it is necessary to limit number of potential overlaps. The idea is to set a cutoff on the relative volume overlap and not on the actual distance between two atoms, since the overlap immediately takes into account the α of the Gaussians. The formula to accept an overlap is defined as follows:

$$\frac{V_{ab}}{V_a + V_b - V_{ab}} \geq 0.03$$

Analysis of the computed volume overlaps indicated that a cutoff of 0.03 and a maximum level of overlap volumes of 6 give a good approximation of the actual hard-sphere volume overlap.

2.3. Molecule-molecule overlap

Given the description of the Gaussian volume of two molecules (`A` and `B`) it is possible to compute the overlap as follows:

$$V_{A,B} = \sum_{a \in A, b \in B} V_{ab} - \sum_{a,b \in A, c \in B} V_{abc} - \sum_{a \in A, b,c \in B} V_{abc} + \sum_{a,b \in A, c,d \in B} V_{abcd} - \ldots$$

Again, the procedure starts with all possible atom-atom overlaps and works from there further through all possible overlaps. An overlap is possible if the Gaussians that make up the overlap exist. For instance, to compute the overlap V_{abc}, this is only possible if V_{ab}, V_{ac}, and V_{bc} exist. By eliminating overlaps as early as possible it reduces significantly the number of combinations to process.

One important remark to be made at this point is that the overlap of a molecule with itself will not give the same value as the total volume of the molecule. This is due to the nature of the Gaussian overlap. The parameterization of the Gaussians was chosen such that the overlap of an atom with itself produces the hard sphere volume. However, overlapping the same atom more than twice with itself will give a deviating value of the hard sphere volume. Therefore, it is important to compute the self-overlap of a molecule to get the right reference value.

2.4. Optimal alignment

2.4.1. Initial orientation

The actual alignment procedure needs to start from good initial guesses. At first, the two molecules are aligned such that their respective center of mass and their principal axes coincide.

If there are *N* atoms and overlap Gaussians, the center of mass is computed as

$$\left(m_x, m_y, m_z\right) = \frac{1}{V}\left(\sum_i x_i V_i, \sum_i y_i V_i, \sum_i z_i V_i\right).$$

The principal axes of the molecule are the eigenvectors of the singular value decomposition of mass matrix of higher-order moments:

©2007-2010 Silicos NV

$$M = \frac{1}{V}\begin{bmatrix} \sum_i V_i x_i x_i & \sum_i V_i x_i y_i & \sum_i V_i x_i z_i \\ \sum_i V_i y_i x_i & \sum_i V_i y_i y_i & \sum_i V_i y_i z_i \\ \sum_i V_i z_i x_i & \sum_i V_i z_i y_i & \sum_i V_i z_i z_i \end{bmatrix}.$$

Translating the molecules such that the centers of mass coincides, allows rewriting the alignment as a rigid-body rotation. To facilitate the rotational optimization, quaternion algebra is introduced[2]. The formulas for the volume overlap are rewritten as:

$$V_{ab} = pp\sqrt{\left(\frac{\pi}{\alpha+\beta}\right)^3}\exp\left(-\frac{\alpha\beta}{\alpha+\beta}q'Aq\right),$$

where *q* is a unit quaternion representing the rotation and *A* is a matrix that solely depends on the position of the two centers of Gaussians *a* and *b*.

2.4.2. Gradient ascent

Since the quaternion of the rotation should have a unit norm, the optimization problem is actually a constrained one with elliptic constraints. The procedure used here is based on a procedure outlined by Hasan and Hasan[3]. The optimization problem is formulated as:

Maximize VAB(q) subject to |q|2= 1

The update formula can be rewritten, using Lagrange multipliers, as:

$$q_{t+1} = q_t + \alpha h$$

with:

$$h = \nabla_q V_{AB}(q) - 2\lambda q$$

and:

$$\alpha = -\left(h'(\nabla_q^2 V_{AB}(q) - \lambda)h\right)^{-1} h'h$$

and the lagrange coefficient is:

$$\lambda = q'\nabla_q V_{AB}(q)$$

Computing gradient and Hessian turns out to be rather straightforward. The gradient is:

---

[2] Karney, C.F.F., J. Mol. Graph. Mod., *Quaternions in molecular modeling*, 25(5), 596-604.

[3] Hasan A.A., Hasan M.A., 2003. In Proceedings ICASSP 2003, Constrained gradient descent and line searcg for solving optimization problem with elliptic constraints. pp II.793-796.

©2007-2010 Silicos NV

$$\nabla_q V_{ab}(q) = -2V_{ab}Aq$$

and the Hessian is:

$$\nabla_q^2 V_{ab}(q) = 2V_{ab}(2Aqq'A - A)$$

which only depends on the already computed volume, the matrix A and the quaternion q.

2.4.3. Simulated annealing

The gradient ascent gives mostly satisfactory results; however in a few cases the procedure gets stuck in a local optimum. To overcome this problem an additional simulated annealing[4] approach is also implemented. The overall procedure is outlined in the following scheme:

```
Start from best rotor quaternion found
While ( i < nbr_iter )
{
    T = sqrt(i/ΔT);
    Perturbate rotor quaternion
    Compute volume overlap Vnew
    If ( Vnew > Vbest )
    {
        Vbest = Vnew; update best_rotor;
    }
    If ( exp(-sqrt((Vold-Vnew)/T)) > random(0,1) )
    {
        Vold = Vnew; keep rotor quaternion;
    }
}
return Vbest, best_rotor;
```

The most important parameters of the simulated annealing procedure are the number of iterations and the temperature. The number of iterations is a user-defined parameter that can be set using the '`--addIterations`' option. The temperature is adapted each iterations based on a ΔT, which is set to 1.1. This value for ΔT is deduced from several test runs and gives good optimization properties.

2.5. Alignment scores

2.5.1. Scores

Given the total volume overlap of atoms V_a and the respective self-overlap volumes of reference (V_{ra}, V_{rp}) and database molecule (V_{da}, V_{dp}), several different scores can be computed:

[4] http://en.wikipedia.org/Simulated_annealing

©2007-2010 Silicos NV

- Taninoto : $s = V_a / \left(V_{ra} + V_{da} - V_a \right)$
- Reference Tversky : $s = V_a / \left(0.95 V_{ra} + 0.05 V_{da} \right)$
- Database Tversky : $s = V_a / \left(0.05 V_{ra} + 0.95 V_{da} \right)$

2.5.2. Score file format

The score file is a tab-delimited text file with 8 columns. The first line is a header line containing the tags of the columns. Using the definitions from the previous paragraph, the 8 columns are ordered and labeled as follows:

1. dbName : identifier of database molecule
2. refName : identifier of the reference molecule
3. Piramid:: Tanimoto
4. Piramid::Tversky_Ref
5. Piramid::Tversky_Db
6. Overlap : effective overlap volume of the atoms
7. refVolume : Volume of reference molecule
8. dbVolume : Volume of database molecule

©2007-2010 Silicos NV

3. Usage

3.1. Command line interface

Piramid is run from the command line as follows:

```
> piramid [options]
```

Options can be either required or optional. Specifying no option at all or the '`-h`' or '`--help`' option generates a help message:

```
> piramid -h
```

Using the option '`-v`' or '`--version`' will print the version of the program and the version number of the libraries on which Piramid depends.

The following sections describe in detail the different options.

3.2. Required command line options

3.2.1. Reference molecule

The name of the file containing the reference molecule is specified using the '`-r`' or '`--reference`' option:

```
> piramid -r filename.ext [other_options]
```

or

```
> piramid --reference filename.ext [other_options]
```

The reference file should contain at least one molecule specified as a connection table with 3D coordinates. Only the first molecule in the reference file will be used as a reference. The remainder of the file is neglected. The format of the input file should be one of the input formats recognized by Open Babel. If the filename carries the extension '.gz', the file is assumed to be an gzipped file.

3.2.2. Database molecules

The file or list of files containing the database molecules is specified using the '`-d`' of '`--database`' option:

```
> piramid –d dbase.ext [other options]
```

or

```
> piramid --database dbase.ext [other options]
```

The database file is a molecule file that could be possibly gzipped. The extension of the database file is used to guess the type. The format of the input file should be one of the input formats recognized by Open Babel.

3.3. Output options

There are a two possible ways to represent the results. The results can be appended to the aligned molecules is a moleculae file or as a tab-delimited output file with the scores. At least one of these two options should be specified, otherwise an error message will be printed.

©2007-2010 Silicos NV

3.3.1. Output molecules

The name of the molecule file to which the molecules are written after alignment to the reference molecule is specified using the '-o' or '--out' option:

```
> piramid -o filename.ext [other_options]
```

or

```
> piramid --out filename.ext [other_options]
```

The output file contains first the reference molecule followed by the molecules aligned to the reference molecule. If the '--best' option is used only the *N* best scoring molecules will be reported in the output file otherwise all molecules are written to output file.

Each molecule in the molecule file contains an additional list of properties in which the respective scores are reported. These fields all start with the tag 'Piramid::'.

3.3.2. Score file

The resulting scores are written to a tab-delimited text file with the options '-s' or '--scores':

```
> piramid -s filename [other_options]
```

or

```
> piramid --scores filename [other_options]
```

The score file contains several columns in which different scores are represented. A more detailed description of the different scores is given in paragraph 2.5.

3.4. Optional command line options

3.4.1. List of best scoring molecules

When only the *N* best scoring molecules need to be reported, the '--best' option is used:

```
> piramid --best N [other_options]
```

With this option set, Piramid will process all molecules and keep track of the best scoring ones. The criterion for 'best' scoring is defined using the '--rankBy' option.

3.4.2. No alignment

If the option '--scoreOnly' is provided, the database molecules are not aligned to the reference molecule :

```
> piramid --scoreOnly [other_options]
```

In this case, only the volume overlap of the given poses with the reference molecule is computed and no optimal alignment is pursued.

3.4.3. Ranking

The reported molecules can be selected or ranked based on a specific score. This is done with the '--rankBy' option :

©2007-2010 Silicos NV

```
> piramid --rankBy <TANIMOTO|TVERSKY_DB|TVERSKY_REF> [other_options]
```

The type of score by which molecules are ranked or selected is defined by a code:

- TANIMORO: Taninoto
- TVERSKY_REF: Reference Tversky
- TVERSKY_DB: Database Tversky

The formulas of these scores are given in Section 2.5. The code is not case-sensitive, therefore uppercase or lowercase forms can be used. This option has an effect when the '`--best`' option is selected to determine the *N* best scoring molecules.

The default score by which molecules are selected or sorted is the Tanimoto score (TANIMOTO).

3.4.4. Additional iterations

By default only a gradient ascent step is performed starting from the initial orientations. If the option '`--addIterations`' is specified, an additional *N* steps of a simulated annealing procedure are performed :

```
> piramid --addIterations N [other_options]
```

This option slows down the program, but might give better alignments when the gradient ascent method gets stuck in a local optimum.

3.4.5. Cutoff

To report only molecules with an alignment score higher then a given cutoff value, the '`--cutoff`' options is used:

```
> piramid --cutoff v [other_options]
```

The cutoff value should be between 0 and 1.0. By default the value is set to 0, which means all molecules will be reported.

©2007-2010 Silicos NV

4. Summary

Table 1. Overview of all command line options.

Argument[5]	Default value[6]	Description
General		
`[O] -h, --help`	N/A	Provides a short description of usage.
`[O] –v, --version`	N/A	Provides the version number of the program.
Input		
`[R] -r, --reference`	-	Defines the reference molecule that will be used to screen and/or align a database.
`[R] -d, --dbase`	-	Defines the database that will be screened and/or aligned.
Output		
`[O] -o, --out`	-	The transformed database molecules after aligning them to the reference molecule.
`[O] -s, --scores`	-	Tab-delimited text file with for the best *N* or each molecule the number of corresponding overlap scores.
`[O] --cutOff`	0.0	Minimum score for a structure to be reported.
`[O] --best`	0	Only the *N* best scoring molecules are reported.
`[O] --rankby`	'TANIMOTO'	Define scoring used by `--cutOff` and `--best`.
`[O] --noRef`	N/A	Do not include the reference molecule in the final output file.
Options		
`[O] --scoreOnly`	N/A	Flag to indicate that no alignment should be computed only a score of the given pose is computed.
`[O] --addIterations`	0	Perform *N* additional optimization steps with the simulated annealing procedure.

[5] [O] means that the argument is optional, and [R] indicates that the argument is required. Not specifying a required argument will cause the program to quit.

[6] N/A stands for 'not applicable'.

©2007-2010 Silicos NV

5. Revision history

5.1. Version 1.0

Initial release

©2007-2010 Silicos NV