Finding Many Stable Molecular Arrangements

Conformational Searching with Genetic Algorithms

Evan Curtin

November 30, 2016

University of Illinois at Urbana-Champaign

Outline

1. Background Information

2. The Genetic Algorithm

3. Finding Low Energy Conformers of Dipeptides

4. Concluding Remarks

Background Information

The Problem

Computational methods require knowledge of molecular structure

The Problem

Computational methods require knowledge of molecular structure

⇒ We need to find the lowest energy structure

The Problem

Computational methods require knowledge of molecular structure

$\Rightarrow$ We need to find the lowest energy structure

The potential energy surface (PES) is high dimensional and has many minima

The Problem

Computational methods require knowledge of molecular structure

$\Rightarrow$ We need to find the lowest energy structure

The potential energy surface (PES) is high dimensional and has many minima

$\Rightarrow$ We can't tell for sure if we've found the **global** minimum

The Problem

Computational methods require knowledge of molecular structure

⇒ We need to find the lowest energy structure

The potential energy surface (PES) is high dimensional and has many minima

⇒ We can't tell for sure if we've found the **global** minimum

We may need information about one or more low-energy conformations

The Problem

Computational methods require knowledge of molecular structure

$\Rightarrow$ We need to find the lowest energy structure

The potential energy surface (PES) is high dimensional and has many minima

$\Rightarrow$ We can't tell for sure if we've found the **global** minimum

We may need information about one or more low-energy conformations

$\Rightarrow$ Ok, let's find them all!

Possible Solutions

- Many techniques are well established

Method	Implented in
grid-based	CEASAR, **Open Babel, Confab,** MacroModel, MOE
rule-based	**ALFA, CONFECT,** CORINA, ROTATE, **COSMOS**, OMEGA
population-based	**Balloon, Cyndi**
basin-hopping	**ASE, GMIN, TINKER SCAN**

Supady, A.; Blum, V.; Baldauf, C. J. Chem. Inf. Model. 2015, 55 (11), 23382348.

Possible Solutions

- Many techniques are well established
- None are perfect

Method	Implented in
grid-based	CEASAR, **Open Babel, Confab,** MacroModel, MOE
rule-based	**ALFA, CONFECT,** CORINA, ROTATE, **COSMOS**, OMEGA
population-based	**Balloon, Cyndi**
basin-hopping	**ASE, GMIN, TINKER SCAN**

Supady, A.; Blum, V.; Baldauf, C. J. Chem. Inf. Model. 2015, 55 (11), 23382348.

Desirables

What Algorithmic Properties do we want for conformer search?

1. Accurate energies & Structures, *ab initio* or DFT

Desirables

What Algorithmic Properties do we want for conformer search?

1. Accurate energies & Structures, *ab initio* or DFT
2. Minimize number of geometry optimizations

Desirables

What Algorithmic Properties do we want for conformer search?

1. Accurate energies & Structures, *ab initio* or DFT
2. Minimize number of geometry optimizations
3. Find the entire low energy population of conformations

Desirables

What Algorithmic Properties do we want for conformer search?

1. Accurate energies & Structures, *ab initio* or DFT
2. Minimize number of geometry optimizations
3. Find the entire low energy population of conformations
4. Minimal human input

Desirables

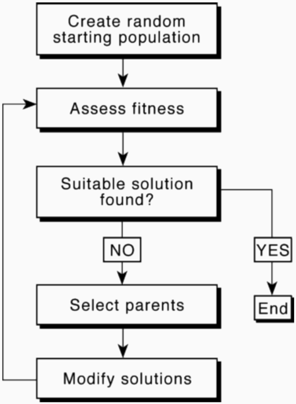
What Algorithmic Properties do we want for conformer search?

1. Accurate energies & Structures, *ab initio* or DFT
2. Minimize number of geometry optimizations
3. Find the entire low energy population of conformations
4. Minimal human input
5. Parallel-Scalable

The Genetic Algorithm

Outline

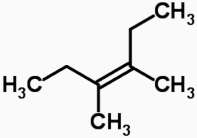
- Inspired by biological evolution
- Evolve a population over generations
- Survival of the fittest
- Requirements:
    - Represent individuals as vector
    - Fitness function
- $v = \begin{pmatrix} x_1 & y_1 & z_1 & x_2 & y_2 & z_2 & \cdots & x_N & y_N & z_N \end{pmatrix}$
- $F = \frac{E_{\text{max}} - E}{E_{\text{max}} - E_{\text{min}}}$



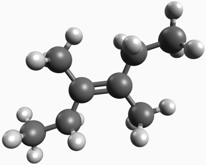
Carwright, H. Reviews in Computational Chemistry 2007 (25), pp 349389.

The Many Representations of a Molecule

2D Image



3D Image



Name

(3Z)-3,4-Dimethyl-3-hexene

SMILES

CCC(C)=C(C)CC

InChI

1S/C8H16/c1-5-7(3)8(4)6-2/h5-6H2,1-4H3/b8-7-

Cartesian Coords

	x	y	z
C	0.90	-0.25	0.02
C	2.35	0.15	-0.17
C	2.91	1.30	-0.67

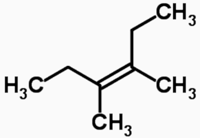
Internal Coords

		r		θ
C				
C	1	1.51		
C	2	1.38	1	131

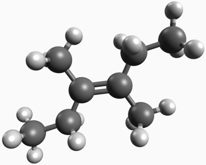
1. Supady, A.; Blum, V.; Baldauf, C. J. Chem. Inf. Model. 2015, 55 (11), 23382348.
2. http://www.chemspider.com/Chemical-Structure.2298795.html

The Many Representations of a Molecule

2D Image



3D Image



Equivalent in theory

Name

(3Z)-3,4-Dimethyl-3-hexene

SMILES

CCC(C)=C(C)CC

InChI

1S/C8H16/c1-5-7(3)8(4)6-2/h5-6H2,1-4H3/b8-7-

Cartesian Coords

	x	y	z
C	0.90	-0.25	0.02
C	2.35	0.15	-0.17
C	2.91	1.30	-0.67

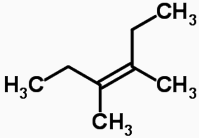
Internal Coords

		r		θ
C				
C	1	1.51		
C	2	1.38	1	131

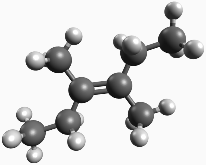
1. Supady, A.; Blum, V.; Baldauf, C. J. Chem. Inf. Model. 2015, 55 (11), 23382348.
2. http://www.chemspider.com/Chemical-Structure.2298795.html

The Many Representations of a Molecule

2D Image



3D Image



Equivalent **in theory**

Name

(3Z)-3,4-Dimethyl-3-hexene

SMILES

CCC(C)=C(C)CC

InChI

1S/C8H16/c1-5-7(3)8(4)6-2/h5-6H2,1-4H3/b8-7-

Cartesian Coords

	x	y	z
C	0.90	-0.25	0.02
C	2.35	0.15	-0.17
C	2.91	1.30	-0.67

Internal Coords

		r		θ
C				
C	1	1.51		
C	2	1.38	1	131

[1] Supady, A.; Blum, V.; Baldauf, C. J. Chem. Inf. Model. 2015, 55 (11), 23382348.
[2] http://www.chemspider.com/Chemical-Structure.2298795.html

Selecting Parents

- Several methods are common

Selecting Parents

- Several methods are common
- Reinforce good characteristics

Selecting Parents

- Several methods are common
- Reinforce good characteristics
- Still give losers a chance

Selecting Parents

- Several methods are common
- Reinforce good characteristics
- Still give losers a chance
- 'Breed' pairs of winners

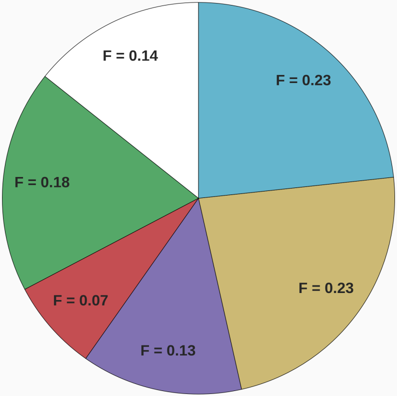
Selecting Parents

- Several methods are common
- Reinforce good characteristics
- Still give losers a chance
- 'Breed' pairs of winners
- 'Roulette Wheel' Method

Sum of Fitness Scores = 1.0

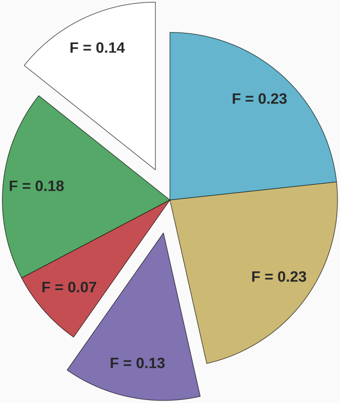
Selecting Parents

- Several methods are common
- Reinforce good characteristics
- Still give losers a chance
- 'Breed' pairs of winners
- 'Roulette Wheel' Method

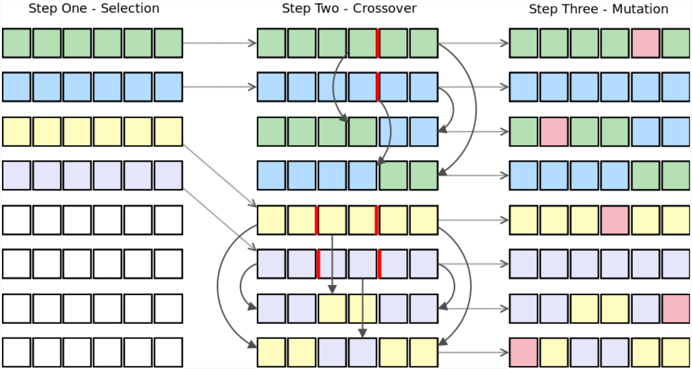


Selecting Parents

- Several methods are common
- Reinforce good characteristics
- Still give losers a chance
- 'Breed' pairs of winners
- 'Roulette Wheel' Method

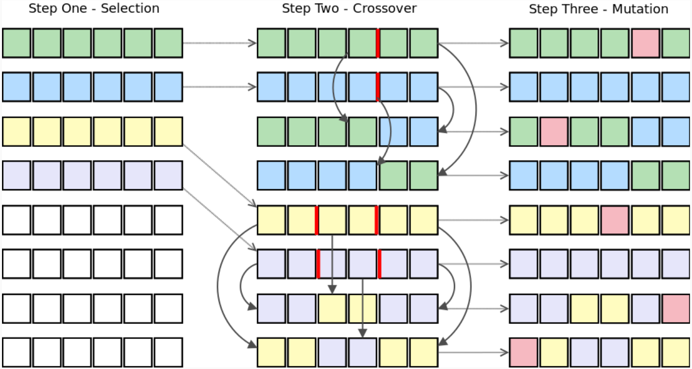


The Next Generation



http://www.turingfinance.com/computational-finance
-updates-ieee-world-congress-computational-intelligence

The Next Generation



Crossover distinguishes this from Monte Carlo

http://www.turingfinance.com/computational-finance-updates-ieee-world-congress-computational-intelligence

The Whole Algorithm

1. Generate N random, sensible geometries

The Whole Algorithm

1. Generate N random, sensible geometries
2. Add each to blacklist

The Whole Algorithm

1. Generate N random, sensible geometries
2. Add each to blacklist
3. Optimize each geometry

The Whole Algorithm

1. Generate N random, sensible geometries
2. Add each to blacklist
3. Optimize each geometry
4. Select Parents

The Whole Algorithm

1. Generate N random, sensible geometries
2. Add each to blacklist
3. Optimize each geometry
4. Select Parents
5. Crossover & Mutate into sensible geometry

The Whole Algorithm

1. Generate N random, sensible geometries
2. Add each to blacklist
3. Optimize each geometry
4. Select Parents
5. Crossover & Mutate into sensible geometry
6. Add Children to population

The Whole Algorithm

1. Generate N random, sensible geometries
2. Add each to blacklist
3. Optimize each geometry
4. Select Parents
5. Crossover & Mutate into sensible geometry
6. Add Children to population
7. Remove High energy individuals

The Whole Algorithm

1. Generate N random, sensible geometries
2. Add each to blacklist
3. Optimize each geometry
4. Select Parents
5. Crossover & Mutate into sensible geometry
6. Add Children to population
7. Remove High energy individuals
8. If converged:

The Whole Algorithm

1. Generate N random, sensible geometries
2. Add each to blacklist
3. Optimize each geometry
4. Select Parents
5. Crossover & Mutate into sensible geometry
6. Add Children to population
7. Remove High energy individuals
8. If converged:
    - Done!

The Whole Algorithm

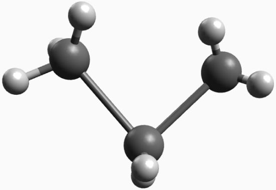
1. Generate N random, sensible geometries
2. Add each to blacklist
3. Optimize each geometry
4. Select Parents
5. Crossover & Mutate into sensible geometry
6. Add Children to population
7. Remove High energy individuals
8. If converged:
    - Done!

   Otherwise:

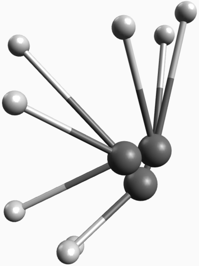
The Whole Algorithm

1. Generate N random, sensible geometries
2. Add each to blacklist
3. Optimize each geometry
4. Select Parents
5. Crossover & Mutate into sensible geometry
6. Add Children to population
7. Remove High energy individuals
8. If converged:
    - Done!

   Otherwise:
    - Go to 2



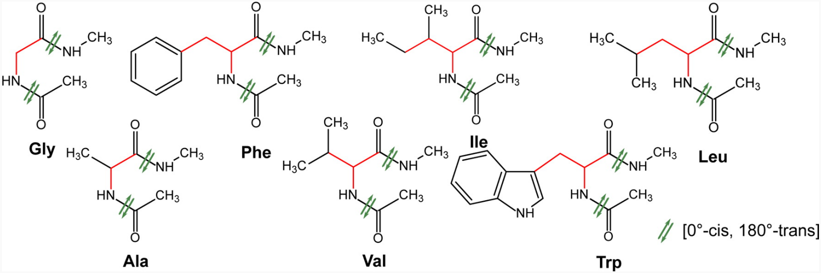
sensible



utter nonsense

Finding Low Energy Conformers of Dipeptides

Dipeptide Structures



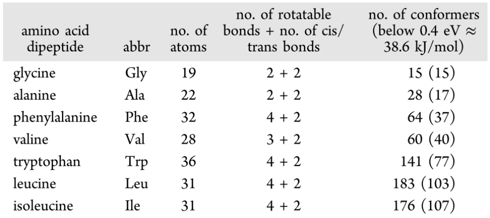
Red = Rotatable bonds

$\rightleftharpoons$ = Cis/Trans Bonds

Supady, A.; Blum, V.; Baldauf, C. J. Chem. Inf. Model. 2015, 55 (11), 23382348.

Combinatorics

- GA beats other methods if space is large
- Space gets large **fast**

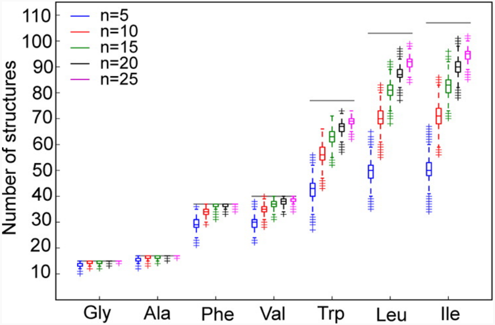


amino acid dipeptide	abbr	no. of atoms	no. of rotatable bonds + no. of cis/trans bonds	no. of conformers (below 0.4 eV ≈ 38.6 kJ/mol)
glycine	Gly	19	2 + 2	15 (15)
alanine	Ala	22	2 + 2	28 (17)
phenylalanine	Phe	32	4 + 2	64 (37)
valine	Val	28	3 + 2	60 (40)
tryptophan	Trp	36	4 + 2	141 (77)
leucine	Leu	31	4 + 2	183 (103)
isoleucine	Ile	31	4 + 2	176 (107)

Supady, A.; Blum, V.; Baldauf, C. J. Chem. Inf. Model. 2015, 55 (11), 23382348.

Coverage

- Smaller systems are reliably sampled
- As # of conformers increases, miss more and more
- Is there a pattern to what is missed?

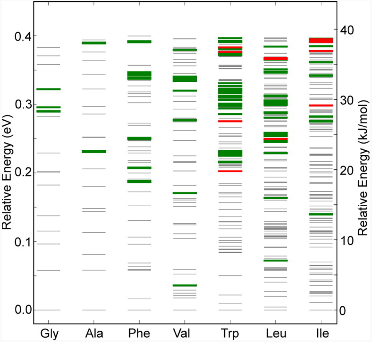


Proportion of conformers found with incrasing runs of the GA

Supady, A.; Blum, V.; Baldauf, C. J. Chem. Inf. Model. 2015, 55 (11), 23382348.

Coverage

- Most misses are very high energy

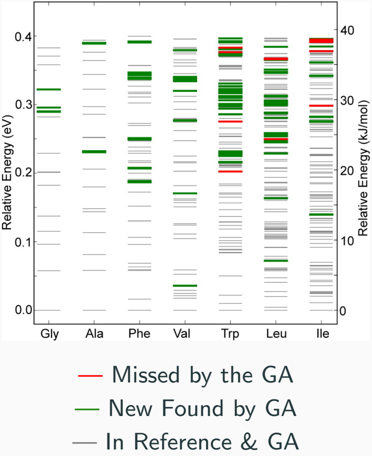


— Missed by the GA
— New Found by GA
— In Reference & GA

Supady, A.; Blum, V.; Baldauf, C. J. Chem. Inf. Model. 2015, 55 (11), 23382348.

Coverage

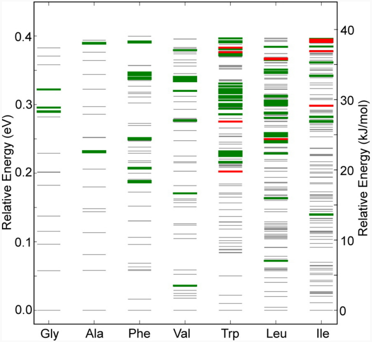
- Most misses are very high energy
- Algorithm favors low energy areas of the space



Supady, A.; Blum, V.; Baldauf, C. J. Chem. Inf. Model. 2015, 55 (11), 23382348.

Coverage

- Most misses are very high energy
- Algorithm favors low energy areas of the space
- Features low in energy are favored and recombined



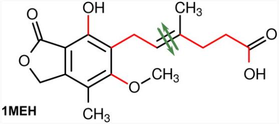
— Missed by the GA

— New Found by GA

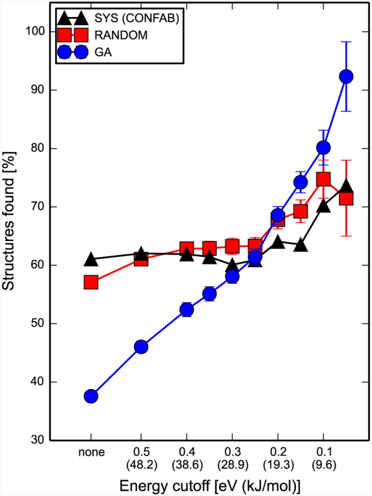
— In Reference & GA

Supady, A.; Blum, V.; Baldauf, C. J. Chem. Inf. Model. 2015, 55 (11), 23382348.

Energy Cutoff

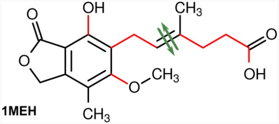


Mycophenolic Acid



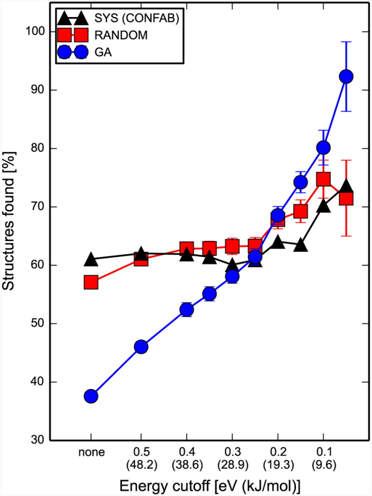
Supady, A.; Blum, V.; Baldauf, C. J. Chem. Inf. Model. 2015, 55 (11), 23382348.

Energy Cutoff



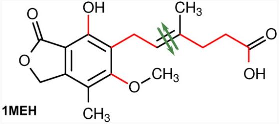
Mycophenolic Acid

- GA is more sensitive to energy cutoff



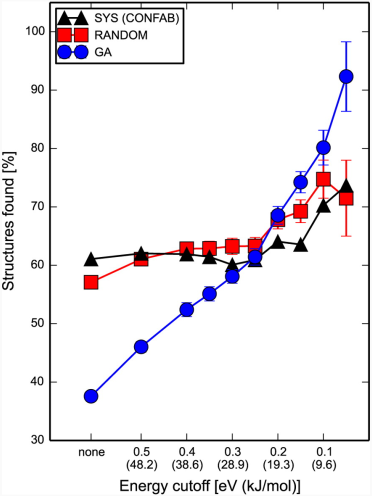
Supady, A.; Blum, V.; Baldauf, C. J. Chem. Inf. Model. 2015, 55 (11), 23382348.

Energy Cutoff



Mycophenolic Acid

- GA is more sensitive to energy cutoff
- For finding low energy ensemble, GA outperforms purely stochastic/deterministic method



Supady, A.; Blum, V.; Baldauf, C. J. Chem. Inf. Model. 2015, 55 (11), 23382348.

Concluding Remarks

Concluding Remarks

- Finding all the low energy conformers for a molecule is hard

Supady, A.; Blum, V.; Baldauf, C. J. Chem. Inf. Model. 2015, 55 (11), 23382348.

Concluding Remarks

- Finding all the low energy conformers for a molecule is hard
- In order to get accurate energies/structures, computationally expensive methods should be employed

Supady, A.; Blum, V.; Baldauf, C. J. Chem. Inf. Model. 2015, 55 (11), 23382348.

Concluding Remarks

- Finding all the low energy conformers for a molecule is hard
- In order to get accurate energies/structures, computationally expensive methods should be employed
- These take time, so we want to minimize how many of these we do

Supady, A.; Blum, V.; Baldauf, C. J. Chem. Inf. Model. 2015, 55 (11), 23382348.

Concluding Remarks

- Finding all the low energy conformers for a molecule is hard
- In order to get accurate energies/structures, computationally expensive methods should be employed
- These take time, so we want to minimize how many of these we do
- The Genetic Algorithm provides a framework for a refined global search

Supady, A.; Blum, V.; Baldauf, C. J. Chem. Inf. Model. 2015, 55 (11), 23382348.

Concluding Remarks

- Finding all the low energy conformers for a molecule is hard
- In order to get accurate energies/structures, computationally expensive methods should be employed
- These take time, so we want to minimize how many of these we do
- The Genetic Algorithm provides a framework for a refined global search
- It shines when asked to find a host of low energy solutions

Supady, A.; Blum, V.; Baldauf, C. J. Chem. Inf. Model. 2015, 55 (11), 23382348.

Concluding Remarks

- Finding all the low energy conformers for a molecule is hard
- In order to get accurate energies/structures, computationally expensive methods should be employed
- These take time, so we want to minimize how many of these we do
- The Genetic Algorithm provides a framework for a refined global search
- It shines when asked to find a host of low energy solutions
- GA wrapper can be interfaced with a variety of electronic structure packages(NWChem, ORCA) and is available under the GNU Lesser General Public License at `https://github.com/adrianasupady/fafoom`

Supady, A.; Blum, V.; Baldauf, C. J. Chem. Inf. Model. 2015, 55 (11), 23382348.

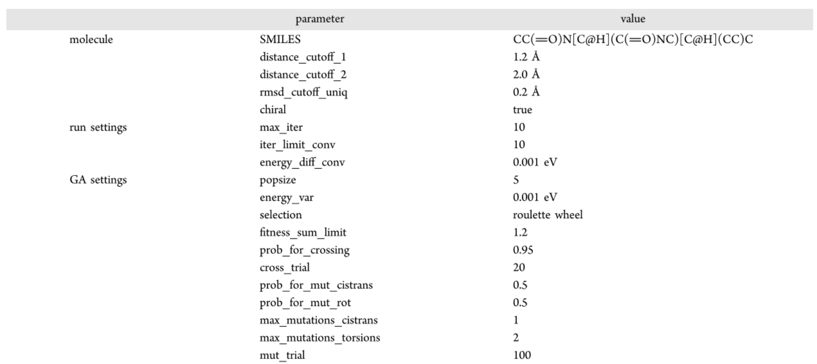
Questions?

Backup slide

- Geometry optimization step makes the algorithm more Lamarckian (Jean Baptiste Larmarck, [1744-1829])

Genetic Algorithm Parameters

Geometry Optimization: DFT PBE + VdW, *tier1* basis in FHI-aims[1]. Convergence at 0.005 eV /Å



	parameter	value
molecule	SMILES	CC(=O)N[C@H](C(=O)NC)[C@H](CC)C
	distance_cutoff_1	1.2 Å
	distance_cutoff_2	2.0 Å
	rmsd_cutoff_uniq	0.2 Å
	chiral	true
run settings	max_iter	10
	iter_limit_conv	10
	energy_diff_conv	0.001 eV
GA settings	popsize	5
	energy_var	0.001 eV
	selection	roulette wheel
	fitness_sum_limit	1.2
	prob_for_crossing	0.95
	cross_trial	20
	prob_for_mut_cistrans	0.5
	prob_for_mut_rot	0.5
	max_mutations_cistrans	1
	max_mutations_torsions	2
	mut_trial	100

GA Parameters for Isoleucine Dipeptide[2]

(1) Blum, V. et.'ál., M. Comput. Phys. Commun. 2009, 180 (11), 21752196.
(2) Supady, A.; Blum, V.; Baldauf, C. J. Chem. Inf. Model. 2015, 55 (11), 23382348.