062_Conformer_Search

April 27, 2022

1 Conformer search

1.1 Outline

- Section **??**
- Section **??**
- Section **??**
- Section **??**
- Section **??**
- Section **??**

1.2 The problem of conformational search

Molecules are not simple 3D objects: they exist as an ensemble of three-dimensional conformations inter- changing over time

Conformational Search: *Locating stationary points in molecular potential energy surfaces (PES)*

i. e. we want to locate the points in which the 4D entity spends most of the time

because the properties are the consequence of a distribution of conformations at any time

Let's consider dihedral degrees of freedom

Systematic approach: generate each possible structure permutating rotatable bonds and relax. Problem: **combinatorial explosion**.
The number of candidate structures grows as

$$\left(\frac{N}{\delta}\right)^{rot.bonds}$$

where *rot. bonds* is the number of rotatable bonds, N is the search domain (i. e. $[-\pi, \pi]$) and δ is the grid resolution.

References 1. N. M. O'Boyle, T. Vandermeersch, C. J. Flynn, A. R. Maguire and G. R. Hutchison, Confab - Systematic generation of diverse low-energy conformers, J Cheminform, 2011, 3, 8.
2. J.-P. Ebejer, G. M. Morris and C. M. Deane, Freely Available Conformer Generation Methods: How Good Are They?, Journal of Chemical Information and Modeling, 2012, 52, 1146–1158.
3. H. H. Avgy-David and H. Senderowitz, Toward Focusing Conformational Ensembles on Bioactive Conformations: A Molecular Mechanics/Quantum Mechanics Study, J. Chem. Inf. Model., 2015, 55, 2154–2167.
4. D. K. Agrafiotis, A. C. Gibbs, F. Zhu, S. Izrailev and E. Martin, Conformational Sampling of

Bioactive Molecules: A Comparative Study, Journal of Chemical Information and Modeling, 2007, 47, 1067–1086.
5. P. C. D. Hawkins, Conformation Generation: The State of the Art, J. Chem. Inf. Model., 2017, 57, 1747–1756.

Stochastic search: can move from one region of the energy surface to a completely different unconnected region in a single step

How we do know if a randomly selected point is relevant? → **Importance sampling**
A regular grid vs a random walk

Monte Carlo[1]: new structure is accepted or rejected based on the Boltzmann criterion:

$$P \propto e^{-\frac{E_{new}-E_{old}}{kT}}$$

[1]Senderowitz, H. & Still, W. C. Sampling potential energy surface of glycyl glycine peptide: Comparison of Metropolis Monte Carlo and stochastic dynamics. Journal of Computational Chemistry 19, 1294–1299 (1998).

In which space should we carry on the exploration? - Cartesian coordinates: simple to manipulate but not efficent for concerted motions - Generalized coordinates: more variable types (bond lengths, valence and dihedral angles) let explore directly topology

Bond lengths and valence angles have single stiff minima well suited for convex optimization.
Rigid rotor: generating new conformers by just torsional changes with unperturbed bond lengths and angles, the latter being eventually relaxed once the conformer has been located.

1.3 A simple example: glycine in gas phase

Eight known stable conformers:

Minimum structures and Transition states: CCSD(T)/CBS + CV electronic energies (kJ/mol) computed at the CCSD(T)/(CBS+CV)MP2 and (B2PLYP-D3BJ/aug-cc-pVTZ) optimized geometries, respectively.

1. A. G. Csaszar, J. Am. Chem. Soc., 1992, 114, 9568–9575.

2. V. Barone, C. Adamo and F. Lelj, J. Chem. Phys., 1995, 102, 364–370.

3. V. Barone, M. Biczysko, J. Bloino and C. Puzzarini, Phys. Chem. Chem. Phys., 2013, 15, 1358–1363.

4. V. Barone, M. Biczysko, J. Bloino and C. Puzzarini, Phys. Chem. Chem. Phys., 2013, 15, 10094.

2D PES of Glycine:

A simple MC search procedure:

1.3.1 Computational details:

- MC + full geometry optimization.[1]
- B3LYP/6-31+G(d) w. Empirical Dispersion.

- Systematic search would include 1728 points (without considering noise).
- Manipulate cartesian/internal coordinates with Parson's method.[2]

References: 1. Li, Z. & Scheraga, H. A. Monte Carlo-minimization approach to the multiple-minima problem in protein folding. Proceedings of the National Academy of Sciences 84, 6611–6615 (1987). 2. J. Parsons, J. B. Holmes, J. M. Rojas, J. Tsai and C. E. M. Strauss, Practical conversion from torsion space to Cartesian space forin silico protein synthesis, Journal of Computational Chemistry, 2005, 26, 1063–1068.

1.3.2 Results:

- Low acceptance ratio:
- Minima 1-6 (sorted from the **G**lobal **E**nergy **M**inimum) found within **200** iterations.
- But more than **1000** (at 430 K) are needed to detect structures VII and VIII.

How can we reduce the computational cost?

1. Improve the search efficency
2. Reduce the cost per step

Biased MC variant: after being stuck for more than an fixed amount of steps restart from the *highest energy structure* sampled so far. Marginally efficent does not recover structure **VIIIn**.

We need a more efficent exploring algorithm. More on that later.

1.4 Strike a balance: heavy scheme vs cheap scheme

1. Exploit the good performance of low level semi-empirical methods or MM methods in generating correct geometries but less accurate energies.

2. Use a QM/QM' scheme → perform the geometry optimization at low level and evaluate energy higher level.

3. Use a method applicable to the whole periodic table without the necessity of new training → semiempiricals.

Let's try again the glycine benchmark with the DFTBA[1] and PM7 methods[2]:

1. D. Porezag, Th. Frauenheim, Th. Köhler, G. Seifert and R. Kaschner, Construction of tight-binding-like potentials on the basis of density-functional theory: Application to carbon, Phys. Rev. B, 1995, 51, 12947–12957.
2. J. J. P. Stewart, Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters, J Mol Model, 2013, 19, 1–32.

Cheap scheme: 1. Contrained geometry optmization (a specific point in the dihedral grid) with low level method *then single point energy* 2. Filter sampled structures: keep only those within a given threshold from the GEM (20 kJ/mol) 3. Reoptimize these structures after a redundancy check at DFT level (B3LYP/6-31+g(d)-D3)).

Minima	ΔE_{DFT}	ΔE_{DFTBA}	ΔE_{PM7}
Ip_{ttt}	0.0	0.0	0.0
IIn_{ccc}	2.9	8.5	5.6
$IIIp_{tct}$	5.9	4.4	6.7
IVn_{gtt}	7.5	2.5	4.2
Vn_{gtc}	12.0	19.1	2.1
VIp_{ttc}	24.1	14.3	19.3
$VIIp_{tcc}$	2.9	29.7	30.1
$VIIIn_{gtc}$	32.5	31.5	24.1

Let's try a more complex example: Threonine in gas phase

- Six rotatable bonds.
- Complete systematic search at 30° resolution: 12^5 points.
- Szidarovszky et al:[1] 56 unique structures.
- Authors performed 7776 full DTF optimizations B3LYP/6-311++G** refined with 1000+ MP2 calculations
- Seven of them identified with MW spectroscopy[2]

References: 1. Szidarovszky, T., Czakó, G. & Császár, A. G. Conformers of gaseous threonine. Molecular Physics 107, 761–775 (2009). 2. J. L. Alonso, C. Pe´rez, M. Eugenia Sanz, J. C. Lo´pez and S. Blanco, Seven conformers of l-threonine in the gas phase: a LA-MB-FTMW study, Phys. Chem. Chem. Phys., 2009, 11, 617–627

The conformational search was performed (@400K) for 5000 iterations making random changes to just 3 out of 5 dihedrals, and evaluating the trial candidates. Restart if stuck: 100 steps (2% of planned steps)

Method	%acceptance	Q1	Q3	ΔE_{max}	# found	RMSE
DTBA	9.22	13.12	20.44	41.11	39	10.8
PM7	5.53	9.05	17.69	44.60	26	10.6
XTB[1]	7.16	9.86	17.90	79.90	27	9.1

1. C. Bannwarth, S. Ehlert, and S. Grimme, "GFN2-XTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions," J. Chem. Theory Comput. 15(3), 1652–1671 (2019).

- Q1, Q3: average energies of the first and third quartile.
- Energies in kJ/mol
- Complete data sets includes 56 structures
- RMSE: Root Mean Square Error calculated wrt to nearest neighbour in reference

Distribution of structures:

B3LYP/6-31+G(d)-D3 geometries of the seven lowest-energy minima of L-Threonine. The relative energies and percentage probability of occurrence (in parenthesis) estimated at 400 K are reported. Heavy atom RMSD against corresponding MP2/6-311++G(d,p) geometries was less than 0.06 Å

1.4.1 Can MC be used for real world systems?

Ferrocene system studued with chiral spectroscopy (VCD)[1]. Very sensitive: small changes in conformation can completely change the shape.
No Molecular Mechanics FF available and 3 rotatable bonds -> perform search with MC.[2]

1. Ravutsov, M.; Dobrikov, G. M.; Dangalov, M.; Nikolova, R.; Dimitrov, V.; Mazzeo, G.; Longhi, G.; Abbate, S.; Paoloni, L.; Fusè, M.; Barone, V. 1,2-Disubstituted Planar Chiral Ferrocene Derivatives from Sulfonamide-Directed Ortho -Lithiation: Synthesis, Absolute Configuration, and Chiroptical Properties. Organometallics 2021, 40 (5), 578–590.

Use PM7/B3LYP and PM7+*a posteriori* refinement and calculate spectra with final 4 lowest lying states

Final spectra[1] shows the importance of including every structure

1. Chandramouli, B., Del Galdo, S., Fusè, M., Barone, V. & Mancini, G. Two-level stochastic search of low-energy conformers for molecular spectroscopy: implementation and validation of MM and QM models. Phys. Chem. Chem. Phys. 10.1039.C9CP03557E (2019). doi:10.1039/C9CP03557E

1.5 Island model $(\lambda + \mu)$ EA

$(\lambda + \mu)$ EA The standard GA creates a new population at eah step. In the $(\lambda + \mu)$ version a population of μ parents generate λ children and then selection shrink back the population to μ

Avoid clashes

With long flexible chains it is possible to create clashes with arbitrary interpolation (during crossover) or generation of new structures (during mutation).

Interpolation is done rotating in steps toward the max value until a clash is detected. Mutation is tried for a fixed number of steps. Using perception heuristics[1] we check that the adiancency matrix in the molecular graph does not change.

1. Lazzari, F.; Salvadori, A.; Mancini, G.; Barone, V. Molecular Perception for Visualization and Computation: The Proxima Library. J. Chem. Inf. Model. 2020, acs.jcim.0c00076. https://doi.org/10.1021/acs.jcim.0c00076

Test 1

- Run EA with 20 starting structures for 50 generations.
- Selection pressure=0.5
- Cross over and mutation probability=0.5, 0.2 (parents), 0.4 children
- Total 1194 constrained semiempirical optimizations.
- Found 39 unique structures and *all* experimental ones

Island model variant Take a bigger population ($n = 100$) for fewer steps ($g = 20$) and divide in equal size *islands* ($n_i = 4$); selection takes place within islands. Every few generation a new operator, **Migration** overwrites the *less similar* of the next island with its best one in a Round Robin fashion using the L_∞. Islands works on subsets of dihedrals.

Best replica (PM7/B3LYP-D3) 550 Full DFT optimization needed to recover all but 1 (number 21 at 19.45 kJ/mol above GEM)

1.5.1 Further benchmarks: serine and cysteine

Serine: five rotatable bonds. Complete systematic search at 30° resolution: 248832 points. He and Allen et al[1]: 15552 HF/6-31G* calculations, 89 refined at MP2/cc-pVTZ. Found 85 unique structures. 1200 calculationf from EA to find all but three (1600 all but one).

Cysteine: five rotatable bonds. Wilke et al.[2]: 11 664 structures @ HF/3-21G yielded 71 structures refined at MP2/cc-PVTZ. 67 of 71 structures found after 1200 calculations

1. He, K. & Allen, W. D. Conformers of Gaseous Serine. J. Chem. Theory Comput. 12, 3571–3582 (2016).
2. Wilke, J. J., Lind, M. C., Schaefer, H. F., Császár, A. G. & Allen, W. D. Conformers of Gaseous Cysteine. J. Chem. Theory Comput. 5, 1511–1523 (2009)

Cysteine structures retrieved in searches with a RMSD thresholds of 0.2 Å and 0.125 Å

Cysteine structures retrieved in searches with a RMSD thresholds of 0.2 Å and 0.125 Å

1.5.2 Rhodium complexes

Two rhodium coordination complexes described in ref. [1]. Carry out PES exploration to calculate VCD spectra. The role of two explicit solvent molecules (CD3CN) filling the axial positions of the octahedral coordination shell of rhodium atoms is also investigated.

1. G. Szilvágyi, Z. Majer, E. Vass, and M. Hollósi, "Conformational studies on chiral rhodium complexes by ECD and VCD spectroscopy," Chirality 23(4), 294–299 (2011)

Procedure: - Even single point DFT energy is too expensive → exploration carried out only at PM7 level.
- Filter out structures above 25 kJ/mol from the GEM then cluster remaining ones with *Partition around medoids*. Feature space of dihedrals and L_∞ distance. - Find the best number of clusters with multiple scores and pick the centroid of each.

Before 2nd step 14-24 structures for each complex, futher reduced by full optimization at DFT. Final VCD calculated on the four lowest lying.

1.6 Beyond dihedral angles

While covering a very wide chemical space, dihedral angles are not the only possible coordinate. For instance, what about a coordination complex?

Silver aqua ion test case:

LAXS and EXAFS data indicate a linear coordination for the Ag^+ ion which is in agreement with CPMD simulations

How to perform this search?

Devise new operators to work in cartesian space:

Cross over is carried out interpolating coordinates using the *Simulated Binary Crossover* (SBX)[1]:
- extract a uniformly distributed random number $\mu \in [0, 1]$ and fix a parameter η which measures

the closeness of offspring to parents then calculate β:

$$\mu \in [0, 0.5] \rightarrow \beta = 2\mu^{\frac{1}{\eta+1}}$$

$$\mu \in [0.5, 1] \rightarrow \beta = \frac{1}{2}(1-\mu)^{\frac{1}{\eta+1}}$$

then interpolate the parent's coordinates:

$$C_1 = 0.5[(1+\beta)P_1 - (1-\beta)P_2]$$
$$C_2 = 0.5[(1+\beta)P_2 + (1-\beta)P_1]$$

the operation is carried out on nearest neighbour fragments.

1. Llanio-Trujillo, J. L.; Marques, J. M. C.; Pereira, F. B. An Evolutionary Algorithm for the Global Optimization of Molecular Clusters: Application to Water, Benzene, and Benzene Cation. The Journal of Physical Chemistry A 2011, 115 (11), 2130–2138.

Mutation is performed selecting among six different types of motions:: - *rattle*, which applies a gaussian displacement to atomic coordinates of the selected fragment; - *rotate* applies a random quaternion to a fragment c.o.m. - *orbit* applies a random quaternion to fragment in the whole complex c.o.m - *swap* which swaps the center of mass of the fragment with another one - *mirror* which reflects the coordinates of a fragment through a random plane

From the previous results design a minimal system and start searches from *unphysical coordinates*. Details: - $Ag(H_2O)_6^+$ - try both two stage (semiempirical opt. + DFT s. p.) and one stage (DFT opt.) - use mean field to simulate the rest of the solvent

Since in this system flexibility is quite high, using distances, angles or RMSD is not a very efficent distance measure. We used the Ultrafast Shape Recognition method of Ballester and coworkers to run the cluster analysis.[1]

1. Ballester, P. J.; Westwood, I.; Laurieri, N.; Sim, E.; Richards, W. G. Prospective Virtual Screening with Ultrafast Shape Recognition: The Identification of Novel Inhibitors of Arylamine N -Acetyltransferases. J. R. Soc. Interface 2010, 7 (43), 335–342.

Searches carried out with semiempiricals (PM7, HFM1[1] and GNF-xTB) tend to produce antiprismic or 4+2 coordination quite early (HFM centroids shown):

1. V. K. Prasad, A. Otero-de-la-Roza, and G. A. DiLabio, "Atom-centered potentials with dispersion-corrected minimal-basis-set Hartree–Fock: An efficient and accurate computational approach for large molecular systems," J. Chem. Theory Comput. 14(2), 726–738 (2018)

B3LYP however finds the following structure located +0.57 kJ/mol from the GEM of the search:

To further investigate the source of linearity and the limits of the used electronic methods we reoptimized one cluster centroid obtained with a semiempirical employing a modified SDD Effective Core Potentials with just one electron:

Note that this is yet another specific case: there is no need to use only **either** dihedrals **or** quaternion coordinates, they can be mixed with the proper manipulation tools.

More general: as long as there is a mapping from one set of coordinates to another and constraints can be expressed on both the coordinates used in the GA **do need not** to be the same used in the Electronic Structure Code (**ESC**), which can choose the ones it likes.

Flexible: we can mix different ESCs under the same representation at EA level.

More flexible: Proxima could take care of giving the features at EA level. Search and reduction/filtering carried out with the same representation.

1.7 The importance of seeding

1.7.1 and one last case study

Metaheuristics, including the present $(\lambda + \mu)$ IM are developed to avoid to early convergence. The results shown so far (and in the literature) show that they perform quite well on chemical Potential Energy Surfaces.

Question 1: disregarding the specific algorithm used, how it is possibile to further improve perfomance? Can some existing knowledge be used to create populations?
Question 2: how well the $(\lambda + \mu)$ IM performs with respect to other methods, e.g. accellerated sampling?

New case study: aspartic acid in gas phase. Investigated with microwave spectroscopy[1] and accelerated sampling[2]. Calculations identifed 19 distinct structures after B3LYP/6-311G++ and MP2 calculations

1. M. E. Sanz, J. C. López, and J. L. Alonso, Phys. Chem. Chem. Phys. 12, 3573 (2010).
2. Comitani, F.; Rossi, K.; Ceriotti, M.; Sanz, M. E.; Molteni, C. The Journal of Chemical Physics 2017, 146 (14), 145102.

Case study details: perform four replica searchs of Asp with: - $(\lambda + \mu)$ IM same settings as for Thr, Ser, Cys - $(\lambda + \mu)$ IM with optimized initial population - CREST[1] as accelerated sampling method - use always GFN-xTB for energy and geometry optimization

Optimized initial population
In the space of the six dihedrals we generate, instead of a uniform distribution, a **Latin Hypercube**[1] to ensure that the initial population is as diverse as possible. A Latin Hypercube distribution in 2D is like having a rooks in chessboard that cannot threaten each one:

1. McKay, M.D.; Beckman, R.J.; Conover, W.J. (May 1979). "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code". Technometrics. American Statistical Association. 21 (2): 239–245

#

Results with respect to the original MP2 structures:

run	convergence	not found	max RMSD (Ang)	min DE (kJ/mol)
EA/LH/run1	2250	9, 17	0.215	> 20
EA/LH/run2	2850	**0**	NA	NA
EA/LH/run3	2250	**0**	NA	NA
EA/LH/run4	3650	9, 14	0.246	> 20
EA/run1	1750	9, 17	0.352	> 20

run	convergence	not found	max RMSD (Ang)	min DE (kJ/mol)
EA/run2	2950	**0**	NA	NA
EA/run3	3450	2	0.248	> 20
EA/run4	3150	**0**	NA	NA
crest/run1	107\2500	15	0.311	> 20
crest/run2	107\2500	15	0.326	> 20
crest/run3	108\2520	15	0.313	> 20
crest/run4	103\2560	15	0.310	> 20

- EA average convergence: 2825 +/- 645, **no miss**
- EA/LH average convergence: 2750 +/- 574, **no miss**
- CREST average convergence: 2520 +/- ? (not directly available), **one miss**

Note that CREST uses gradients …

Searches started LH find low energy structures earlier:

Further tests carried out with other semiempirical methods (EA only) do not find all the 19 structures:

GFN-xTB understimate barriers which is actually advantageous:

We tried to measure this difference using the **Gini coefficent**[1], an indicator of the asymmetry of values in a distribution (often used in economics). It is defined as

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n^2 \langle x \rangle}$$

1. Gini, Corrado (1936). "On the Measure of Concentration with Special Reference to Income and Statistics", Colorado College Publication, General Series No. 208, 73–79.

Gini coeffincient of $-\Delta E$

Search	GC
DFTBA/1	0.431
DFTBA/2	0.422
DFTBA/3	0.447
DFTBA/4	0.493
xTB/1	0.304
xTB/2	0.308
xTB/3	0.300
xTB/4	0.296

DFTBA is more elitist than xTB!

The Latin Hypercube is a black box sampling method. But stable forms in the conformer space are not uniformly sparse → exploit perception and domain knowledge to: - check against databases for fragments - generate *sensible* structures