Global Optimization Algorithms for Chemical Applications
Algorithms, Showcases and Other Projects

Johannes M. Dieterich

Institute for Physical Chemistry
Georg-August-University Göttingen
Germany

2012
THIS IS A STRIPPED-DOWN VERSION.
The search space

General problem

- The search space typically has various minima of different fitness.
- Only fewest (normally just one) are of *optimal* fitness.
- In many cases, the number of minima increases exponentially with search space dimensionality.
- *NPOs* (NP optimization problems):
  - knapsack (packing problem)
  - travelling salesman (shortest path problem)
Exponential scaling of search spaces: Should a chemist care?

The Lennard-Jones example

- LJ clusters seem to exhibit exponential (exp($x^2$)) scaling of the number of minima w.r.t. cluster size\textsuperscript{1,2}
- Clusters described by two-body forces are \textit{NP-hard} optimization problems.\textsuperscript{3}

\textsuperscript{2}\text{LJ}_{100} is considered to have $1 \cdot 10^{40}$ local minima.
(Efficient) Exploration of the search space: Techniques

MC/MD-based methods

- *Simulated annealing (SA)*\(^4\): Crystal growth.
- *Monte Carlo with minimizations (MCM)*\(^5\): Basin hopping.

Nature inspired algorithms

- *Genetic algorithms (GA)*\(^7\): Darwinian/Lamarckian evolution.
- *Evolutionary strategies (ES)*\(^8\): Mutation-focused evolution.
- *Particle-Swarm optimization (PSO)*\(^9\): Model movements in a group.

---

Genetic Algorithms for Cluster Optimization

Status

- Highly optimized, system-specific implementations\(^{10}\)
- Typically only support for a few levels of theory
- Maximal support for binary systems\(^{11}\)

Remaining Challenges

- More efficient and general algorithms
- Extensible, futureproof framework
- Efficient and stable parallelization schemes
- Support for arbitrary levels of theory
- Support for arbitrary systems
  - arbitrary systems
  - arbitrary degrees of freedom

\(^{10}\) e.g. Hartke working group: similar yet different versions for TIP4P clusters, doped TIP4P clusters, TTM2F clusters, clathrates,...

\(^{11}\) see e.g. R. L. Johnston \textit{et al} and R. Ferrando \textit{et al}
Genetic algorithms in a nutshell

(Genetic) Operations

- selection/mating
- crossover
- mutation
- (sanity check)
- (optional magic)
- local optimization

Maintaining/updating the population

- diversity check
- add to the population
Genetic algorithms in a nutshell

(Genetic) Operations

- selection/mating
- crossover
- mutation
- (sanity check)
- (optional magic)
- local optimization

Maintaining/updating the population

- diversity check
- add to the population

(b) crossover
Genetic algorithms in a nutshell

(Genetic) Operations

- selection/mating
- crossover
- mutation
- (sanity check)
- (optional magic)
- local optimization

Maintaining/updating the population

- diversity check
- add to the population
Genetic algorithms in a nutshell

(Genetic) Operations

- selection/mating
- crossover
- mutation
- (sanity check)
- (optional magic)
- local optimization

Maintaining/updating the population

- diversity check
- add to the population

(d) collision (i.a.)

(e) dissociation (i.a.)
Genetic algorithms in a nutshell

(Genetic) Operations

- selection/mating
- crossover
- mutation
- (sanity check)
- (optional magic)
- local optimization

Maintaining/updating the population

- diversity check
- add to the population
Genetic algorithms in a nutshell

(Genetic) Operations

- selection/mating
- crossover
- mutation
- (sanity check)
- (optional magic)
- local optimization

Maintaining/updating the population

- diversity check
- add to the population
Genetic algorithms in a nutshell

(Genetic) Operations

- selection/mating
- crossover
- mutation
- (sanity check)
- (optional magic)
- local optimization

Maintaining/updating the population

- diversity check
  - add to the population
Genetic algorithms in a nutshell

(Genetic) Operations

- selection/mating
- crossover
- mutation
- (sanity check)
- (optional magic)
- local optimization

Maintaining/updating the population

- diversity check
- add to the population
So, what is the *magic*?

Options

- *directed mutation*: move a building block to a better position (requires system knowledge)\(^{12}\)
- *explicit exchange*: (randomly) exchange building blocks for a better mixing in highly-mixed systems\(^{13}\)
- *hybrid algorithms*: couple GA/MCM for better search space exploration and exploitation\(^{14}\)

**GA/MCM**

---

\(^{12}\) Applications to water clusters by B. Hartke *et al.*, currently in development for OGOLEM

\(^{13}\) See e.g. R. L. Johnston, *Dalton Trans.* 4193 (2003).

Differences in search space exploration

Exploration and exploitation differences

- **GA-based methods**: good long-range exploration, good short-range exploitation (with minimizations)
- **MC/MD-based methods**: good mid-range exploration and exploitation, good short-range exploitation (with minimizations)
Benchmarking genetic algorithms: Ackley’s function

Function definition

\[ f(\vec{x}) = -20 \cdot \exp \left[ -0.2 \sqrt{\frac{1}{n} \sum_{i=1}^{n} x_i^2} \right] - \exp \left[ \frac{1}{n} \sum_{i=1}^{n} \cos(2\pi x_i) \right] + 20 + e^1 \] (1)

with the global minimum \( x_i = 0.0 \) in the \( n \)-dimensional space.\(^{15}\)

Benchmorking genetic algorithms: Ackley’s function II

Conclusions

- The scaling is both w/ and w/o local optimization excellent for all tested crossover operators ($O(N^{1.1})$ and $O(N^{1.0})$).
- The benchmark function can be easily solved even for 1000 dimensions.
- Standard benchmark functions are too simple for GA-based global optimization techniques.\textsuperscript{16}

\textsuperscript{16} J. M. Dieterich and B. Hartke, \textit{in preparation}.
Hypersurfaces for cluster optimization

Requirements

- exact description (to the sub-kcal regime)
- fast energy and gradient evaluations (since typically multiple hundreds of thousands are needed)
- well-behaving (no discontinuities, distinct gradients)

(Partial) Solutions

- specialized force fields: fast but potentially inaccurate
- *ab initio* methods: exact and (very) slow
- combinations of the two
- any other methods (semiempirics,...) depending on the system
OGOLEM methods and interfaces

Interfaces

- **Tinker** and **OpenBabel**: standard force fields
- **Amber, Gromacs** and **NAMD**: (custom) force fields for larger biological systems
- **DFTB+**: tight-binding methods
- **MNDO** and **MOPAC**: (special) semiempirical methods
- **Orca**: semiempirics, DFT and *ab initio*
- **CPMD** and **CP2K**: plane wave DFT with periodic boundaries
- **MOLPRO**: high(er)-level *ab initio* methods

Internal implementations

- **Standard LJ FF**: employing Lorentz-Berthelot mixing rules
- **High-level LJ FF**: dedicated parameters for each pair, arbitrary power expression
- **Gupta FF**: two-body FF for metals
- **Amber FF**: reparametrizable
- **UFF**
- **GGFF**: general Gaussian-based two- and three-body FF
OGOLEM: Global optimization of chemical problems

Cluster structures

- arbitrary levels of theory
- arbitrary systems with arbitrary DoF
- first implementation of a $N$-phenotype X-over for $N > 2$
- support for systems in environments (on surfaces, in pockets)

Parametrization

- (system-specific) potential parametrization against reference data
- currently focussed on energy-structure dependence for fitness

Molecular design

- discrete optimization problem
- design of photo-switchable molecules\(^\text{17}\)
- tuning of substuents for tuning of excitation energies
- excitation energies using FOCI-AM1 (Mopac) or TDDFT (Orca)

Kanamycin A dimers

Methods

- Global optimization based on AMBER:GAFFv10 employing genotype operators.
- Reoptimization using MOPAC:PM3.
- Final structure selection based on MOLPRO:DF-LMP2/cc-pVDZ.
- Structures seem to support the preference of sodium over potassium also seen in experiment.\textsuperscript{18}

Sodium-doped water clusters I: Techniques

Experimental requests

- relatively exact description of \((\text{H}_2\text{O})_n\text{Na} (n<7)\) clusters
- coverage of the *relevant* conformational space
- spectral prediction (IR)

Methods

- phenotype global optimization
- level of theory in the optimization: B3LYP/6-31+g** (Orca)
- post-processing: ZPVE correction based on harmonic frequencies
- post-processing: DF-LCCDS(T0)/aug-cc-p(C)VTZ single points
- automatic assignment of groups based on number of Na-O contacts and hydrogen bonds
- calculation of averaged spectra
Sodium-doped water clusters II

\[(H_2O)_3Na\]

- allows for a reliable assignment of structural patterns
- may explain temperature dependence of experimental spectra\(^\text{19}\)
- level of theory becomes unfeasible for \((H_2O)_XNa\) with \(X > 7\)

scaTTM3F: A modern TTM3-F

Project

• global optimization of water clusters (bigger than 34 units)
• using TTM3-F as a backend
• Java2JNI2Fortran is nasty (SMP impossible)

Status

• purely Scala-based implementation (scaTTM3F)
• Åland (energy decomposition based globopt)
• niching
• directed mutation TBD
Water clusters: Some common TTM2-F/TTM3-F structures

(a) 3  (b) 4  (c) 5  (d) 6  (e) 7  
(f) 8  (g) 9  (h) 10  (i) 11  (j) 12  
(k) 13  (l) 14  (m) 15  (n) 16
The workflow

\[ V_{tot}(r_2) = V_1(R - r_2) + V_2(r_2) + V_3(R - r_2), \quad R = 16.34 \text{Å} \]  

- constrained BP86/def2-SVP optimizations (rotated planes for BF4)
- DF-MP2/cc-pVTZ single points
- GGFF[5] fitting
- Is this really reasonable?
Clever opt: Simplification rules! II

(a) BF<sub>4</sub> pot.

(b) Cl pot.

(c) (BF<sub>4</sub>)<sub>3</sub>

(d) (BF<sub>4</sub>)<sub>2</sub>Cl

(e) (BF<sub>4</sub>)Cl<sub>2</sub>
Sn(III)/Ge(III) Hydrides

Results

- BP86/def2-SVP Sn: SD-ECP optimizations and frequencies
- NBO analysis: Sn-H: 89.3% p-character, Ge-H: 84.8%
- Lone pair: Sn: 79.4% s-character, Ge: 70.9%
- 2nd order perturbation energy NBO analysis: N lone pair donation into metal p orbital (50 kcal/mol Sn, 35 kcal/mol Ge)

---

Sn(III)/Ge(III) Hydrides: DFT-D?

What about DFT-D?

- local geometry around the metal similar to BP86
- isopropyl geometry changes (methyl moiety points towards metal), difference to experiment
- DF-MP2/cc-pVTZ favours the BP results (by 2.9 for Sn and 0.2 kcal/mol for Ge)
Atomdroid: Mobile computational chemistry

User interface

- Molecular viewer
- Molecular builder
- Molecule manipulation/analysis
- PDB downloader
- Bluetooth sharing

Computational methods

- UFF energy and gradient\textsuperscript{21}
- L-BFGS\textsuperscript{22}, NEWUOA and Powell minimizations
- Monte Carlo
- Monte Carlo with Minimizations (MCM)\textsuperscript{23}


Benchmarks: Is it more than just a toy?

Prejudice

- Mobile platforms are (and will always be) just for toys not for real applications.
- Hardware is not (and will never be) fast enough for real computations.

Benchmark systems

(c) MeOH  (d) Coffein  (e) Cryptand 222  (f) Neurokinin A  (g) alpha RgIA

Hardware: PoV Mobii Tablet Tegra2 w/ Android 3.0 ROM
Benchmarks: Results

<table>
<thead>
<tr>
<th>System</th>
<th>No. of atoms</th>
<th>energy [ms]</th>
<th>energy+gradient [ms]</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6</td>
<td>0.10±0.01</td>
<td>0.18±0.02</td>
</tr>
<tr>
<td>B</td>
<td>24</td>
<td>0.86±0.01</td>
<td>1.48±0.02</td>
</tr>
<tr>
<td>C</td>
<td>63</td>
<td>3.54±0.02</td>
<td>5.48±0.02</td>
</tr>
<tr>
<td>D</td>
<td>155</td>
<td>17.40±0.09</td>
<td>26.47±0.07</td>
</tr>
<tr>
<td>E</td>
<td>201</td>
<td>28.74±0.03</td>
<td>43.56±0.02</td>
</tr>
</tbody>
</table>
Summary and Outlook

- general framework for global optimization
- exploitation of new hardware platforms: Android, CUDA (not in this talk)
- development of new parallelization schemes
- implementation and parametrization of classical force field expressions as well as novel, flexible ones
- cooperations with experiment (Kanamycin A, doped water clusters)
- development of novel GA-based algorithms
- ...
Acknowledgements

The group:
Ricardo Mata
João Oliveira
Jonas Feldt
Thorsten Stolper
Milica Andrejic

group-leader
biosystems
mobile computing
opts and regions
QM/MM

Collaborations:
Bernd Hartke
Sascha Frick
Thomas Zeuch
Svenja Fischmann
Dietmar Stalke
Guido Clever
Mark Feyand

water clusters
hydration shells in reactions
cluster experiments
internship
inorganic compounds I
inorganic compounds II
inorganic compounds III

Bachelors:
Sebastian Gerke (Göttingen), Maximilian Konrad (Göttingen), Sebastian Wille (Göttingen), Marvin Kammler (Göttingen)