Evolutionary Algorithm for the Bonded Molecular Conformation Problem

Bennett Gale

August 2018

Contents

1 Introduction

The model molecular energy minimisation problem involves the geometric configuration of a chain of atoms in two dimensions, each connected by rigid bonds of unit length, such that the total potential energy yielded by the sum of individual Lennard-Jones potentials is globally minimised. The total potential energy of the cluster is given by

$$V(r) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(r_{ij}^{-12} - 2r_{ij}^{-6} \right)$$

Where r is the euclidean distance between a pair of atoms and N is the number of atoms in the chain.

The energy minimisation problem is historically difficult because of the exponential relationship between the locally stable structures and the number of atoms, N, in the molecule. That is, as the size of the cluster increases, the number of local minima increases exponentially such that when N approaches 100, there exist more than 10^{140} minima [1]. It is therefore necessary to employ sophisticated search algorithms in order to make larger input sizes attainable.

Section 2 will discuss some of the research that has already taken place in this area. Section 3 will detail the algorithm proposed by this paper.

2 Literature Review

Genetic algorithms have been applied [2], [3] successfully in order to extend the range of input size N that is reachable through structural optimisation. This population based approach emulates the Darwinian process of natural selection, and works by performing crossover operations on pairs of candidate configurations with selection probabilities based on their fitness, producing new child nodes which contain genetic material from both parents. With every crossover operation there is a small chance of subsequent random mutation of the child configurations, the purpose of which is to promote genetic diversity within the population and prevent premature convergence. Over many generations the population converges towards a global minimum as each configuration communicates information about the global search space through the crossover operation. This method has been successful in producing results that were previously unseen in literature for certain input sizes [3].

Judson *et al.* compare the performance of GAs with simulated annealing, a probabilistic global optimisation technique in which Monte Carlo steps are taken and new states are accepted if they are fitter, or with a probability based on a function of a temperature value [2]. When applied to the molecular conformation problem, they found that the GA method and the SA method, both applied as global meta-heuristics in conjunction with gradient-based local optimisers (conjugate gradient), similarly outperformed purely stochastic methods [2]. Judson *et al.* classify points on the model energy surface as being in one of

three states: knotted and of high energy, unknotted and of low but not minimal energy, and unknotted with globally minimised energy [2]. They applied both GA and SA in order to make the transition from the first state to the second state before applying local optimisation to descend to the bottom of the basin.

Genetic operators for the two dimensional bonded Lennard-Jones problem have since been further developed with successful results [3], proving the GA approach to be a viable global optimisation technique. Pullan [3] proposes three genetic crossover methods and four mutation operators and applies these in a parallel genetic algorithm to find global minima for most values of $N, 2 \leq N \leq 55$, as well as a minimum for $N = 61$ which was previously unseen in literature.

This two stage hybrid method has been successful, but Judson *et al.* also suggest the possibility of adding a third optimisation stage which is intermediate between the global and local gradient methods [2]. It is thought that a non-gradient optimisation method such as the simplex method may provide a means of stepping over barriers into possibly deeper minima surrounding those found by the global method in the first phase.

More recently, the Conformational Space Annealing Method [4] has been developed as a global optimisation algorithm that has produced very good results when applied to the problem of an unbound cluster of atoms with Lennard-Jones relationships. This algorithm combines concepts of Monte Carlo with minimisation, genetic algorithms, and simulated annealing. While this technique has not been applied to the bonded variant of the Lennard-Jones conformation problem as discussed here, it demonstrates that a hybrid approach combining SA and evolutionary population based algorithms may be a viable technique in a similar problem domain.

3 Proposed Algorithm

As discussed in section 2, global meta-heuristics such as the genetic algorithm, have been used successfully as a means of navigating the global search space and avoiding local minima that are caused by knots in the configurations [2], [3]. These methods use the global optimiser to find unknotted potential candidate configurations and then refine them using gradient-based methods. Judson *et al.* have suggested a three stage system in which a non-gradient optimiser such as the simplex method may be used as an intermediate stage between the global and local-gradient optimisation in order to traverse barriers into potentially deeper basins [2].

As simulated annealing has been found to be successful when combined with evolutionary algorithms in similar problem domains [4], this paper will explore its use as an intermediate non-gradient method such that the proposed algorithm has three levels: global meta-heuristic (GA), intermediate non-gradient optimisation (SA) and local gradient-based optimisation (BFGS). The performance of this three stage optimisation method will be compared with a purely evolutionary algorithm without the intermediate step.

3.1 Genetic Encoding

Encoding each molecule as a chain of two-dimensional Cartesian coordinates would require constraints to be enforced so that the relative orientations and unit length bonds are maintained. In order to avoid this, therefore, molecules will be encoded as sequences of $N-2$ angles, where N is the number of atoms in the molecule and each angle, α_i represents the rotation of the segment following atom a_i. This encoding scheme means that Cartesian coordinates must be calculated at each Lennard-Jones potential computation as it requires euclidean distances between pairs of atoms to be computed, but results in much simpler crossover and mutation operations.

3.2 Selection & Crossover Methods

The population is represented as a large pool of size $n_{pop} \times 2$ where n_{pop} is the user-provided population size. This pool contains both the current population and the generated children in a single array such that it can be easily sorted based on fitness.

3.2.1 Crossover

A single-point crossover operator is applied as follows: A single angle index a_i is selected at random. Angles to the left of this point in the first parent form the left segment of the child, and the right segment of the child is similarly formed from the atoms to the right of the second parent.

Partial L-BFGS Finally, α_i is locally optimised by applying BFGS with this single angle as its only optimisation parameter, with the rest of the angles fixed. Energy and gradients for use in the BFGS algorithm are computed using the entire molecule, while only the subset is optimised, leaving the rest of the angles unchanged. This optimisation of the single crossover point effectively locally optimises the relative orientation of the two connected segments. Crossover operators with multiple crossover-points were not found to be beneficial during experimentation and were even detrimental in some cases. As such, they have been excluded from this study. The focus is primarily on evaluating the use of annealing as a mutation operator and therefore the crossover operation has been kept simple.

3.2.2 Selection

Pairs of parent configurations to be crossed over are selected by choosing randomly from the fittest half of the population. This is repeated n_{pop} times so that the number of children created is equal to the population size and forms the next generation. Selecting from the fitter half of the population ensures that information about the global search space is retained through multiple generations by providing bias towards states that are closer to the global minimum.

At each generation the entire pool is sorted so that the n_{pop} fittest states, including children and fitter parents from the current generation move to the top half of the pool. This forms the new generation's population, and implicitly provides a form of elitism by allowing states from the previous generation to remain in the population if they are fit enough. After sorting, the second half of the pool contains discarded states which can be later overwritten with new children during subsequent crossover stages.

3.3 Mutation

The mutation mechanism is where the intermediate non-gradient optimisation algorithm will be incorporated. In traditional genetic algorithms, mutation has a small chance of occurring and is random, resulting in some states which are less fit in the hopes that this will later lead to a deeper basin in the search space. By incorporating simulated annealing at the mutation stage it is hoped that the configuration will be dragged in a downward direction while also providing enough stochastic properties to allow boundaries between minima to be traversed, which is the goal of any mutation operator.

3.3.1 Mutation Rate

Mutation generally has a low chance of occurring so as not to corrupt fit configurations too frequently and lose information provided by the global search. It was found, however, that with a low mutation rate the search tended to converge prematurely due to an overly homogeneous population. In order to overcome this, the chance of mutation for each configuration is based on its similarity to its direct neighbour. That is, each state is mutated if its fitness is within one unit of that of the previous state.

Additionally, if a state's energy is positive it is guaranteed to be mutated. This has the effect of reducing the number of knotted configurations in the population, as the mutation operators tend to guide them towards a negative energy state.

3.3.2 Mutation Operators

Unravelling This mutation operator is based on one described by Pullan [3] in which A random angle index α_i is selected in the molecule and the distance between this point and each endpoint is compared. The angles between the selected angle and the closest end forms a new segment of angles. For each angle in the selected segment, a very small (close to zero) random number is chosen and the angle is set to that value. This results in one end of the molecule being 'unravelled' into a straight line, which effectively removes any knots or kinks within that segment.The segment is later 'rewound' through subsequent local optimisation.

Guided Stochastic Mutation A random angle is selected and the current fitness and the current value of this angle is stored.The selected angle is set to a random value between $-\pi$ and π and the fitness is recomputed. If the new fitness is less than the previous fitness, revert the angle to its previous state. This process is then repeated until the maximum number of iterations is reached. This limit is important because after many generations it becomes increasingly difficult to find a random change in an angle that does not give lower energy, which would otherwise result in an infinite loop.

Naive Mutation

Simulated Annealing Mutation A relatively brief annealing schedule is used due to the fact that during every iteration an energy computation is required. The goal at this intermediate search algorithm is not to find the global minimum, but merely to provide a more localised search for deeper basins in the vicinity of the current configuration; it is hoped that if the GA search has found a local minimum in close proximity to the global minimum, the SA stage will take large enough steps in order to climb over the boundary into the neighbouring basin where it can be finally optimised by the gradient search in the next stage. Therefore, an annealing schedule that prioritises speed over optimality will be chosen. The initial temperature T will be set to 100 and decrement by one every iteration until it reaches 0.

During annealing, random angles are selected and modified by random amounts. If the change produces an improved state, it is always accepted. If the change produces a worsened state, it is accepted with a probability of $1 - e^{-\Delta E/Tk}$ such that both the current 'temperature' as well as the degree to which the change will worsen the configuration, determine whether the change is to be accepted. ΔE represents the change in energy that will occur if the new configuration is accepted, T represents the current temperature, and k is used to scale probability curve. A value of $k = 0.03$ was chosen as this provides a reasonable probability curve for energy deltas of 1 and below.

3.4 Local Optimisation

3.5 GA Procedure and Parameters

4 Results

5 Discussion

6 Appendix

References

[1] D. Daven, N. Tit, J. Morris, and K. Ho, "Structural optimization of lennard-jones clusters by a genetic algorithm," *Chemical Physics Letters*, vol. 256, no. 1, pp. 195–200, 1996, ISSN: 0009-2614. DOI: https://doi.org/10.1016/0009-2614(96)00406-X. [Online]. Available: http://www.sciencedirect.com/science/article/pii/000926149600406X.

[2] J. R. S., C. M. E., M. J. C., H. A., and G. D., "Do intelligent configuration search techniques outperform random search for large molecules?" *International Journal of Quantum Chemistry*, vol. 44, no. 2, pp. 277–290, DOI: 10.1002/qua.560440214. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qua.560440214. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/qua.560440214.

[3] W. Pullan, "Genetic operators for a two-dimensional bonded molecular model," *Computers & Chemistry*, vol. 22, no. 4, pp. 331–338, 1998, ISSN: 0097-8485. DOI: https://doi.org/10.1016/S0097-8485(97)00068-5. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0097848597000685.

[4] J. Lee, I.-H. Lee, and J. Lee, "Unbiased global optimization of lennard-jones clusters for n¡=201 using the conformational space annealing method," *Phys. Rev. Lett.*, vol. 91, p. 080 201, 8 Aug. 2003. DOI: 10.1103/PhysRevLett.91.080201. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.91.080201.