Exploring Systems of Reprodution through Modification of a Genetic Algorithm aimed at Optimizing Molecular Geometry

Madeline W. Elder & Anna A. Zhitnitsky

University of British Columbia, Vancouver, British Columbia, Canada

Abstract

The main objective of this study was to compare the efficiency of various evolutionary systems in maximizing the fitness of the population upon which they act. This was accomplished through the development of a genetic algorithm to optimize the chemical structure of carbon dioxide (CO2), which has a precisely known solution. This allowed for the comparison of efficacy of different versions of the algorithm, each based on a different evolutionary strategy. Each breeding system represents a distinct approach to reproduction, generating unique evolutionary curves; however, the population fitnesses converge toward the optimized solution at approximately the same time. It is concluded that for the given problem all versions are practical, but for more complex problems the path taken to optimization might make one method preferable.

Introduction

Genetic algorithms (GAs) provide a means of solving complex optimization problems. The process is analogous to natural selection acting on a population to induce evolution, including concepts such as fitness, mutation, and crossover (Mitchell, 1998). GAs are used across industries to solve optimization problems in everything from shipping routes and gene expression analysis to investment strategies and chemical reaction pathways (Ross & Corne, 1994). When the algorithm is first initiated, a population of candidate solutions is randomly generated, each with their own genome, which contains all parameters necessary to describe the phenotype. These individuals are then reproduced in an iterative process, producing generations of solutions, with each individual assigned a fitness score. The score is indicative of how close an individual is to the ultimate solution relative to other candidates, and depends on what problem the population is meant to solve (Whitley, 1994). Mutations in the genome may also occur, controlled by a fixed mutation rate. This allows successive generations to score progressively higher in the fitness module, and thus evolve the population towards maximum fitness. The fundamental premise of genetic algorithms is that biological evolution provides a highly effective mode of optimization. This paper aims to extend this analysis beyond random mating, introducing explicit systems of reproduction in order to quantitatively investigate their efficiency outside of their intended biological context.

In this work, a genetic algorithm was written with the aim of optimizing molecular geometry by minimizing the energy of interaction between each of the molecule's constituents. CO2 was used as an exemplar, as it can be modeled with a relatively low number of parameters, which translates into a genome of a manageable size. This project could have been conducted on any optimization problem, but this one was chosen because its precise solution is known. Similar molecular geometry optimization problems have been solved with genetic algorithms, particularly those involving organic molecules. One notable example comes from a 1995 paper by Deaven & Ho, wherein the structure of fullerenes is predicted from an initial set of random coordinates, up to the buckyball sphere structure of C_{60}.

Our experiment investigates the evolutionary role of various reproductive strategies observed in the natural world, as well as how they affect the fitness of the populations upon which they act. We aim to provide a quantitative foundation from which to approach these biological questions through the disciplines of

chemistry, mathematics, physics, and computer programming.

Methods

Program Development

Figure 1 illustrates the structure of the genetic algorithm. Because many generations are often needed to find an optimal solution, maximizing efficiency in genetic algorithms is key. We found that the most efficient way to encode information in our program was to represent the population of molecules in a 2D array, each row of which contains sufficient parameters to describe one candidate:

$$\begin{array}{cccc} & 0 & 1 & 2 & 3 \\ 0 & [P_1]_0 & [P_2]_0 & [P_3]_0 & [F_i]_0 \\ 1 & [P_1]_1 & [P_2]_1 & [P_3]_1 & [F_i]_1 \\ & \ldots & & & \end{array} \qquad (1)$$

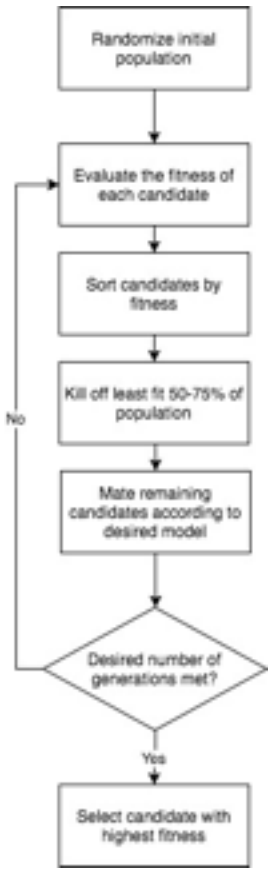

Figure 1. This flowchart shows the general structure of the GA used. The full code can be found in the appendix of this paper.

Parameters 1 and 2 (P1 and P2) represent the C-O bond lengths of the molecule, and were randomly and uniformly drawn from the range 1.0-2.0 Å, a reasonable interval given that the observed bond length is 1.16 Å (Phan et al., 2003). The third parameter represents the bond angle between the two C-O bonds, in the range of 0° to 180°. Finally, the fitness of each candidate was calculated using these same three parameters. The fitness function is based on the electrostatic potential between each charge pair permutation and follows the equation:

$$F = -k\left(\frac{Q_1 \times Q_c}{P_1} + \frac{Q_2 \times Q_c}{P_2} + \frac{Q_1 \times Q_2}{r_3}\right) \qquad (2)$$

$$r_3 = \left(P_1^{\,2} + P_2^{\,2} - 2 \times P_1 \times P_2 \times \cos(P_3)\right)^{\frac{1}{2}}$$

Where $k = 9 \times 10^9$ N (Coulomb's constant), Q1, Q2, and Qc are the partial charges (d+ and d-) on the two oxygen atoms and the carbon atom. The aim is to minimize the net energy of interaction and thus maximize the fitness score. The negative sign ensures the function is strictly non-negative. The combination of the fitness and sort functions inherently eliminates physically impossible phenotypes produced by not assigning them a fitness score. An example of such a structure is when Parameter 3, the O-C-O bond angle, is zero, such that the oxygen atoms lie on top of one another. This renders the third term of the fitness function undefined, represented in Python as NaN (not a number).

We wrote a mutation function into each mating function to maintain genetic diversity. If a mutation is unfavourable, natural selection will prevent its widespread uptake; otherwise, mutations accelerate the process of evolution. In the absence of mutations, a population risks stagnation - where the population converges on the same non-optimal solution (Johnston, 2003). We used a method analogous to a roulette wheel to incorporate static mutations into our algorithm. Static mutations result in the replacement of a gene with a random new allele (Johnston, 2003). In the mutation roulette, a random number between 0 and 1 is generated for each gene of each candidate. If this number is below the predefined mutation rate, the gene is randomly mutated. For instance, if the mutation rate were set to 0.01, probabilistically, 1% of all genes of all candidates would be mutated per

generation.

Upon the creation of a new generation, the candidates are sorted in decreasing order of fitness using an external NumPy sort function. This allows for the straightforward removal of the bottom-most portion of the population, to be replaced with offspring of the fitter candidates.

Methods

Simulation of Biological Systems

In this section, we will compare each biological system modeled: Alpha, League Monogamy, and Polygamy. In the Alpha model, the top candidate of each generation (index 0) mates with each of the other surviving individuals. This is analogous to breeding systems in wolves and lions (Mech, 1999). The rationale behind this system of evolution is that the alpha individual has the best genes. Thus, the quickest way to improve the fitness of the population as a whole is to spread these genes across the entire subsequent generation.

In the League Monogamy model, the bottom 50% of candidates are removed per generation. Then, potential candidates pair with candidates of similar fitness, producing two offspring. This simulates standard human mating, where status often determines reproductive patterns, and the number of children produced is tending towards replacement (Schmitt, 2005).

To model r-selected polygamy, candidates pair several times per generation, producing one offspring per pairing. Additionally, 75% of each generation is removed to allow for more offspring. As in the real world, the evolutionary benefit of polygamy is that it increases the diversity in the next generation, as each individual can potentially have a unique set of parents (Nutting 1891). Such methods of evolution are often found in r-selected species, such as plants and insects, which must be capable of adapting to drastic environmental changes by reproducing at high rates.

Results

Each evolutionary model was run for 100 generations, after which the highest fitness score was recorded. This was considered to be the optimized solution.

The fitness score to which each final solution was compared was calculated using the same function with experimentally determined data: $P_1=P_2=1.16Å$, $P_3=180°$ (Phan et al., 2003). Table 1 summarizes the solution given by each model, as well as relative fitness scores.

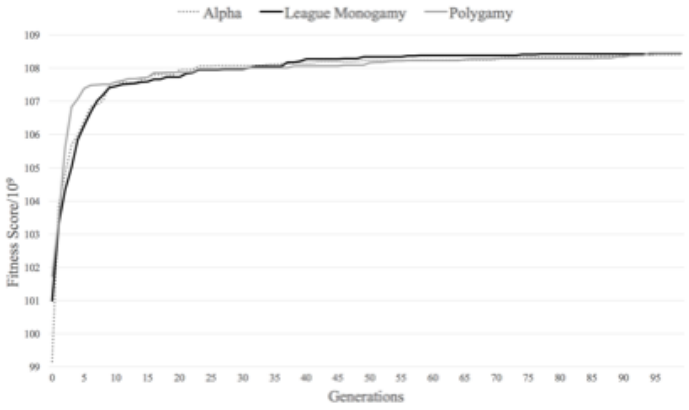

Figure 2. Graph of the highest scoring individual of each generation following the three simulated evolutionary models. The algorithm was run for 100 generations with a fixed mutation rate of 0.1. Standard deviation for each data point is not shown for the sake of visibility; however, the standard deviation obtained for the final fitness scores of the top candidate of each model can be found in Table 1.

Table 1. Optimized data describing the geometry of a CO2 molecule after 100 generations. The relative fitness is the ratio between the calculated fitness score of the ideal known solution and the fitness of the top optimized solution for each breeding system modelled. Therefore, the most accurate model is the one with a relative fitness closest to 1. All values are averages taken from 10 trials. The highest fitness score of each generation is plotted in Figure 2.

Model	Bond Length 1 (Å)	Bond Length 2 (Å)	O-C-O Angle (°)	Final Fitness Score	Relative Fitness	σ in Final Fitness Score (%)
Alpha Male	1.1630	1.163	179.071	1.0841E+11	1.0153	0.12
League Monogamy	1.1608	1.160	179.048	1.0844E+11	1.0155	0.15
Polygamy	1.1636	1.162	179.073	1.0845E+11	1.0156	0.18

Figures 2 and 3 summarize the successive evolution of the highest fitness score of each reproductive model.

Discussion

Figure 2 illustrates the evolutionary pathway followed by each of the three reproductive methods. After 100 generations the curves converge on the same optimized solution, indicating a similar overall efficiency.

However, as highlighted in Figure 3, differences between fitness in earlier generations shows that each breeding system has distinct characteristics. For example, r-selected polygamy appears to be initially the most efficient, due to the rapid initial improvement in population fitness. This correlates to the biological rationale for polygamy, that it allows the population to adapt very quickly to drastic changes in the environment, such as starting from an entirely random molecular geometry (Gould & Lewontin, 1979). However, because the 'best' genes do not dominate the gene pool, the polygamy model experiences the longest plateau before reaching the optimized solution. This suggests that it may not be necessary for an r-selected species to reach an idealized genome. By definition, r-selected species thrive in environments of constant changing pressures, where the requirements for the 'perfect' genome can change regularly. This highlights an inherent limitation of this system, one that makes it unfavourable for species in more stable environments, such as humans (Gould & Lewontin, 1979). It is also worth noting that the results of the polygamy model showed the greatest standard deviation; 0.18%, suggesting a general instability in this evolutionary system.

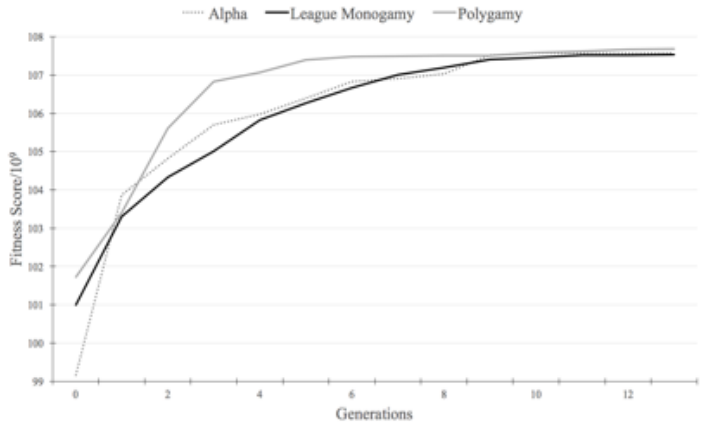

Figure 3. Graph of the highest scoring individual of each generation following the three evolutionary models simulated, for the first 14 generations. As in Figure 2, the standard deviation for each data point is not shown for the sake of visibility.

Similarly, the Alpha Male model displays a sharp initial increase in fitness, as the genes of the top candidate flood the gene pool. However, this results in very low genetic diversity in subsequent generations, risking stagnation. This can be prevented with a sufficiently high mutation rate; however, of the three models, the Alpha model is still most susceptible to achieving false or local maxima, as illustrated in Figure 4. Nevertheless, because of the low genetic diversity, the uncertainty associated with the results of this model is lowest: ±0.12%. In sum, the Alpha model will give the most precise solution, but is best used when an approximate solution is already known. It is important to note that such a method for reproduction is generally only seen in k-selected species, such as lions and wolves, whose environmental pressures remain comparatively stable, meaning they can afford to have relatively low genetic diversity (Huston, 1979).

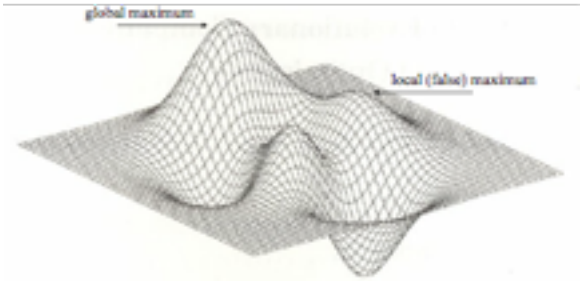

Figure 4. This generalized fitness landscape demonstrates the concept of a local maximum in comparison to a global one. The z-axis represents fitness, while the xy plane represents two parameters. Getting "stuck" at a local maximum is one of the biggest dangers for genetic algorithms, as it can lead to very imprecise and inaccurate results. One of the best defenses against this is high rates of mutation. Adapted from the work by Clegg (2008) without permission (13).

League monogamy is interesting in that it most closely approximates human mating systems. Because it has the most moderate rate of evolution, we posit that it is the most widely applicable model. It is worth noting that this model contains the technically simplest code, which can be useful when approaching a problem that is otherwise complex. However, because of the gradual rate of evolution, it may be inefficient when considering a problem for which the candidate solutions are described by a large number of parameters.

Limitations

We did not design our genetic algorithm to comprehensively simulate biological systems, meaning there are limits to the generalizability of its results. For instance, there was no regulation of successful offspring breeding with their parents. For our purposes, this may be beneficial to the evolution of the population, as both the offspring and parent may be leading

candidates in the race to optimization. However, in biological systems, inbreeding can introduce unfavourable mutations, causing offspring to have lowered fitness or even to die.

Furthermore, in the algorithm proposed here, there was no maximum lifespan, meaning that a high-scoring candidate could maintain its position in the population for many generations, something that is not possible in the biological world.

One of the main limitations of our fitness function was in its physical accuracy. For example, the function did not account for the screening effects of carbon on oxygen, which would have made the 180° structure vastly more favourable. In order to eliminate this effect, a specific if loop was written for candidates with P3=180°, such that their fitness was evaluated according to the following equation:

$$F = -k\left(\frac{Q_1 \times Q_c}{P_1} + \frac{Q_2 \times Q_c}{P_2}\right) \quad (3)$$

This function assumes no oxygen-oxygen interaction due to the screening effects of carbon. Even so, partial screening is not accounted for.

Finally, we note that the final fitness score does not change dramatically in response to large changes in bond angle, which is untrue in experimental settings (Wu & Chern, 1997). This is likely because the derivative of cos(x) is -sin(x), which approaches 0 as x nears the observed bond angle of 180°. This means that with an angle close to 180°, the rate of change of the cos(x) term is very small, even for large changes in x.

Conclusion

This experiment is important not only in the pursuit of the most efficient genetic algorithm, but also to further our understanding of why different species and populations have developed such a diverse range of reproductive strategies: how does each mode affect evolutionary speed? In what ways does this serve the population?

As might be expected, each system of evolution confers advantages and disadvantages to the population upon which it acts. Models reliant on extreme genetic diversity and rapid reproduction, such as the polygamy model, are best suited for determining an approximate solution, as it retains the greatest uncertainty, but takes the least amount of time to run, and explores the largest number of candidate solutions. Conversely, problems that require high precision, and for which the solution is already approximately known, would best be solved by a lower diversity model such as Alpha. We posit that League Monogamy is the most generally applicable model, as it is both relatively efficient and accurate. Furthermore, because it is the simplest to program, the model would be particularly efficient to use on highly complex problems involving many parameters and a sophisticated fitness function. For this particular optimization problem, any evolutionary system could be used to solve the problem in under 100 generations.

Future work could explore the trends discussed above in a more complex optimization problem. This would allow us to see how significant the differences in global maximum uncertainty and fitness function dynamics can be, and whether they might give rise to a preference for one system over the others.

Acknowledgements

We would like to thank Dr. James Berger for his valuable advice on the mechanics of the program and his mentorship throughout our research. We would also like to thank Dr. Elliott Burnell for discussion on how to model molecular geometry, as well as his suggestions for the numerical rates of mutation and death.

References

Clegg, K. (2008). An Introduction to Evolution for Computer Scientists. Deaven, D. M., & Ho, K. M. (1995). Molecular geometry optimization with a genetic algorithm. Physical review letters, 75(2), 288.

Gould, S. J., & Lewontin, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. Proceedings of the Royal Society of London B: Biological Sciences, 205(1161), 581-598.

Huston, M. (1979). A general hypothesis of species diversity. The American Naturalist, 113(1), 81-101.

Johnston, R. L. (2003). Evolving better nanoparticles: Genetic algorithms for optimising cluster geometries. Dalton Transactions, (22), 4193-4207.

Mech, L. D. (1999). Alpha status, dominance, and division of labor in wolf packs. Canadian Journal of Zoology, 77(8), 1196-1203.

Mitchell, M. (1998). An introduction to genetic algorithms. Cambridge: MIT press.

Nutting, C. C. (1891). Some of the causes and results of polygamy among the Pinnipedia. The American Naturalist, 25(290), 103-112.

Phan, A., Doonan, C. J., Uribe-Romo, F. J., Knobler, C. B., O'keeffe, M., & Yaghi, O. M. (2010). Synthesis, structure, and carbon dioxide capture properties of zeolitic imidazolate frameworks. Acc. Chem. Res, 43(1), 58-67.

Ross, P., & Corne, D. (1994). Applications of genetic algorithms. AISB Quaterly on Evolutionary Computation, 89, 23-30.

Schmitt, D. P. (2005). Fundamentals of human mating strategies. The handbook of evolutionary psychology.

Whitley, D. (1994). A genetic algorithm tutorial. Statistics and computing, 4(2), 65-85.

Wu, H. J., & Chern, J. H. (1997). A Remarkable Effect of C– O– C Bond Angle Strain on the Regioselective Double Nucleophilic Substitution of the Acetal Group of Tetraacetal Tetraoxa-Cages and a Novel Hydride Rearrangement of Tetraoxa-Cages. The Journal of organic chemistry, 62(10), 3208-3214.