Toy model protein folding with simulated annealing

Severi Rissanen

April 2019

1 Introduction

Protein folding has been an active area of study for over half a century by now [1]. Understanding the problem is important, as proteins have an important role in the regulation of biologic activities and for example misfolding is a cause of many diseases [2]. Predicting conformations using computational methods thus has much promise in avoiding slow experimental processes and helping us find new protein structures as well as understanding their formation [1].

In this project I studied a computational toy model for protein folding, based on the idea given in Ref. [3]. The model is simplified from the real problem so that there are only two types of amino acids, which are represented essentially as single atoms. The protein conformation corresponding to a shape of some chain of these amino acids should be found by finding the global energy minimum, a claim also known as the thermodynamic hypothesis [4]. The general aim of this project was to gain some understanding of the model and to study methods to find these minimum energy configurations, and in particular a Monte Carlo simulated annealing procedure was programmed and the results compared with a library global optimization routine.

2 Theory

In the model, the amino acids form chains with unit distance links between each of them, and the different possible conformations of the chain are defined by the angles between consecutive links. In the original article, the model was two-dimensional, but I decided to do the computations in three dimensions, which also meant that defining conformations with the angles uniquely was slightly more tricky. Figure 1 illustrates the procedure, which relies on the bend angles θ_i and ϕ_i. $\overrightarrow{AB}$ always points towards the z-direction, while $\overrightarrow{BC}$ is rotated towards the negative y-axis by the angle θ_1, defined between 0 and π. After this, the directions of the θ-rotations are first set to be so that the next bend is rotated clockwise away from the last bend around the vector got from the cross product of the two vectors connecting the last three amino acids in order.

Intuitively, this means that the next bend is rotated in the same direction as the last bend was rotated from the one before it. After this, the ϕ-rotation is applied around the last bend clockwise. In the figure, the position of D is then defined by first calculating $\overrightarrow{AB} \times \overrightarrow{BC} \propto \hat{x}$ and rotating $\overrightarrow{CD}$ around that by θ_2, and then rotating that around $\overrightarrow{BC}$ by ϕ_1. The procedure continues by then rotating $\overrightarrow{DE}$ from $\overrightarrow{CD}$ by θ_3 around $\vec{v}_{\theta_3}$ and rotating the result by ϕ_2 around $\overrightarrow{CD}$.

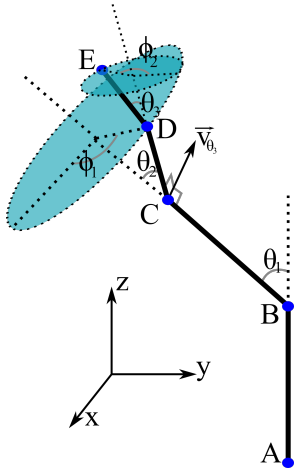

Figure 1: The definition of a conformation of a chain consisting of 5 amino acids with the bend angles θ_i and ϕ_i. The vector $\vec{v}_{\theta_3} \propto \overrightarrow{BC} \times \overrightarrow{CD}$. The letters are labeling the amino acids, and don't refer to the species here.

As in Ref. [3], two types of interaction energies were defined. The so-called backbone bend potentials were set to be

$$V_1(\theta_i) = \frac{1}{4}(1 - \cos\theta_i), \tag{1}$$

the effect of which is to straighten out the chain. Second, Lennard Jones-potentials, set to act between amino acids separated by at least two links, were defined as

$$V_2(r_{ij}, i, j) = 4(\frac{1}{r_{ij}^{12}} + \frac{C_{ij}}{r_{ij}^{6}}) \tag{2}$$

where i and j denote different amino acids within the chain, r_{ij} is the separation between them and C_{ij} is a constant depending on the the species of i and j, which can be either of type A or type B. For two amino acids of type A the constant equals -1, for two B-acids it is -$\frac{1}{2}$ and +1 for acids of differing species. Thus, the potentials can be separated in to strongly attracting, weakly attracting and

strongly repulsing types. The total energy to be minimized was then

$$V_{tot} = \sum_{i=1}^{N-2} V_1(\theta_i) + \sum_{i=1}^{N-2} \sum_{j=i+2}^{N} V_2(r_{ij}, i, j) \tag{3}$$

where N is the total number of acids in the chain.

3 Simulation

In this project, Monte Carlo simulated annealing was used to find minimum energy conformations of chains, that is, a Metropolis algorithm was run with the temperature parameter going towards zero. The variables used in the algorithm were naturally the bend angles $\theta_1, \ldots \theta_{N-2}$ and $\phi_1, \ldots \phi_{N-3}$, where N is the number of acids in the chain. The transition probability distributions were set to be normal distributions centered at the old values and the acceptance probability defined so that the Boltzmann constant was set to 1, resulting in the acceptance rate $e^{-(E_{new}-E_{old})/T}$, where T is the temperature.

In addition to the bend angles, the state of the system was represented in memory with the relative distance vectors $\vec{v}_1, \ldots, \vec{v}_{N-1}$ between consecutive acids in the chain, since this was convenient with regards to finding the coordinates of the acids as the bend angles were varied. The procedure for defining the conformations as outlined in Sec. 2 could then be directly applied using the Rodrigues' formula

$$\mathbf{A} = \mathbf{I}\cos\varphi + (1-\cos\varphi)\hat{e}\hat{e}^T + \begin{bmatrix} 0 & -e_3 & e_2 \\ e_3 & 0 & -e_1 \\ -e_2 & e_1 & 0 \end{bmatrix} \sin\varphi, \tag{4}$$

where $\mathbf{A}$ is a rotation matrix that applies rotations around the vector $\hat{e}$ by the amount φ. This is quite useful especially since a rotation of an individual bend angle during the Metropolis run could be realized simply by applying the given rotation matrix to all relative distance vectors after the bend, optimizing the algorithm quite a bit.

The results of simulated annealing were also compared with a library global optimization routine, which implements the Basinhopping algorithm [5]. Since the algorithm relies on somewhat similar principles and it has been used successfully in similar molecular energy optimization problems [6], it serves as a good benchmark for the created routine.

4 Results

The two optimization routines were compared by running them multiple times on one chain and calculating the proportion of runs that resulted in the lowest energy found on either routine within an accuracy of 0,005. This proportion can be interpreted as a reliability score for obtaining the global minimum, that

at least one run found the global optimum conformation, which isn't of course necessarily true in all cases, but it still serves as a rough metric for comparing the procedures. The two procedures were run on all differing chains of length three, four and five for a hundred, fifty and twenty times, respectively.

An attempt was made to choose good parameters for both algorithms, so that the comparison would be somewhat fair. The state of the system was prepared with Metropolis thermalization before running for both optimizers, and the amount of iterations for the procedures were set so that the time taken was approximately the same. The temperature for annealing was set to start from 0,1 and go to 0 linearly with the Boltzmann constant set to 1, and the temperature setting in the basinhopping set to 1, since these seemed to give good results. The standard deviation for the Metropolis algorithm transition probability was set to $\pi/20$ for the same reasons.

Tables 1 and 2 show the performance of the two algorithms for all differing amino acid chains of length three and four, and the appendices also contain the results for length five in table 3. Only non-symmetric chains were considered, meaning that BAA and AAB were not considered separately, for example. We can see that for simple chains of length three, both algorithms perform quite well an find the optimal conformation almost all the time, but for lengths four and five, some chains chains cause problems, and the calculated reliabilities range all the way from zero to one. It's notable that the simulated annealing routine seems to perform better in most cases, and although there are chains where basinhopping wins, it's not by a particularly large margin.

	BBB	ABA	BBA	BAB	AAA	BAA
Optimal energy	-0.030	-0.66	0.032	-0.030	-0.66	0.032
Annealing reliability	1,0	1,0	1,0	0.99	1,0	1,0
Basinhopping reliability	1,0	1,0	1,0	1,0	1,0	1,0

Table 1: The energy of the optimal configuration from the 100 annealing and basinhopping runs for chains of length 3 and the proportions of runs that reached the correct energy within a deviation of 0,005.

	BBBB	ABBA	BAAB	AAAA	BBBA
Optimal energy	-0,16	-0,036	0,062	-2,32	0,0047
Annealing reliability	0,26	0,02	1,0	1,0	1,0
Basinhopping reliability	0,14	0,08	1,0	1,0	1,0
	BBAB	BBAA	ABAB	ABAA	BAAA
	-0,0008	0,067	-0,65	-1,45	-0,59
	0,98	1,0	0,98	0,98	0,96
	1,0	1,0	1,0	0,98	1,0

Table 2: The energy of the optimal configuration from the 50 annealing and basinhopping runs for chains of length 4 and the proportions of runs that reached the correct energy within a deviation of 0,005.

Figures 2 and 3 illustrate the lowest energy conformations found by simulated annealing. Images for chains of length five are included in the appendices in Fig. 4. It's clear that there are essentially two possible conformations for chains of length three: A bent shape for chains with acids of type A at both ends and a straight line for the others. The attraction between two type-B acids at the ends, in particular, is not enough to win the backbone bend potential and bend the chain. Looking at length-four chains, a few more clearly differing conformations appear, and for length five we find a whole zoo of protein shapes. Qualitatively, a clear pattern is that as the length increases, the proportion of completely straight shapes decreases, from 67% for length three, to 50% for length four and to 20% for length five. The essential ingredient for straight shapes seems to be that there should be less of type A acids than of type B, preventing the strong interactions between type A acids bending the structure, but also not too many B:s, as in the chains with only type B acids. Whether a particular chain has a straight shape or not is not easy to guess, however, and it's interesting to note that even quite small changes, like replacing the type A acid in BBBBA towards the center of the chain as in BBBAB, can result in the straight conformation to change to something completely different.

Figure 5 in the appendix shows a conformation of a chain of length ten, AAABBBBAAA, found using simulated annealing 300 times and taking the lowest-energy result. The result makes intuitive sense, since one would expect that the correct folding puts the ends together and the middle clumped away due to the attractive interactions between acids of the same type and the repulsion between the middle and the ends. Fig. 6 presents a histogram for the optimized energies. There is quite a bit of variation in the results, but the abrupt stop in the distribution at the left side of the figure may indicate that a much lower energy won't be found at least in a same type of conformation as in Fig. 5. Thus, it seems reasonable that the global minimum was found to some degree of accuracy.

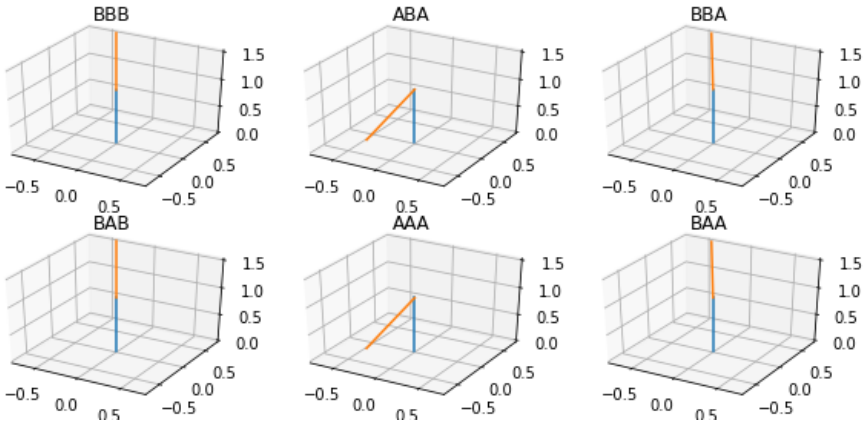

Figure 2: The optimal conformations of the six different chains of length 3 found using simulated annealing. The first acid in the chain always starts at (0,0,0).

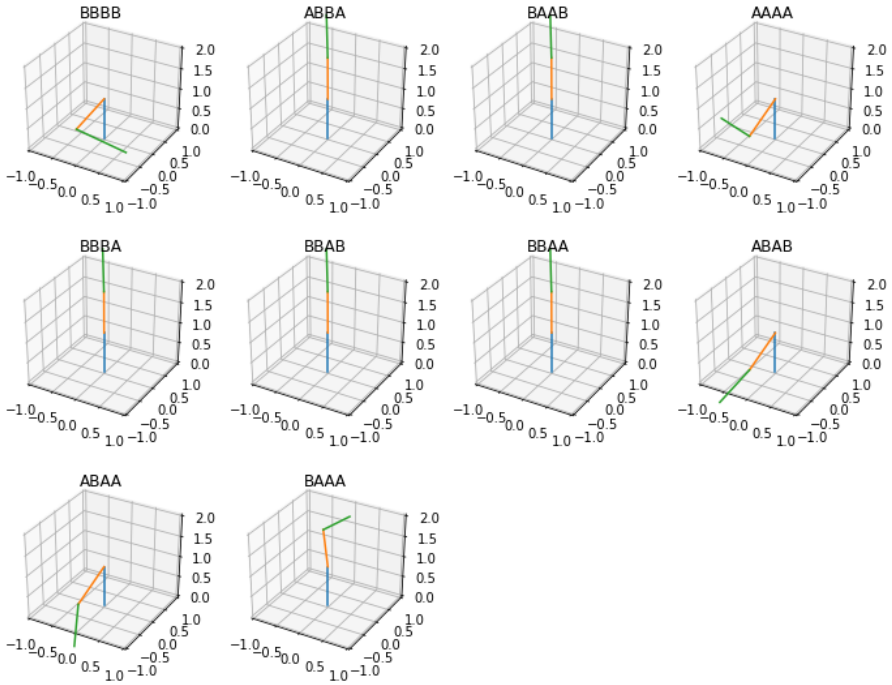

Figure 3: The optimal conformations of the six different chains of length 4 found using simulated annealing. The first acid in the chain always starts at (0,0,0).

5 Conclusions

The aim of this project was to implement a simulated annealing procedure to find minimum energy configurations for toy-model amino acid chains, that is, the corresponding protein conformations. The capability of the method was assessed by comparing its results with a library global optimization routine that uses the basinhopping algorithm, and some qualitative remarks were made with respect to the folded shapes. It was found that the performance of the self-programmed method was at least comparable to the library routine, and possibly even somewhat better. This doesn't necessarily mean that simulated annealing is a better algorithm, since the calculation of the energy was better optimized for it and both methods were given more or less the same time to run, but we can conclude that implementing your own optimization routine can give better results quite easily than an out-of-the-box one in this problem.

There are many ways to improve the implemented method. Much thought wasn't put into the shape or width of the transition probability distribution of the Metropolis algorithm, and better choices surely exist. The distribution could also be different for θ and ϕ -angles. Monitoring the acceptance rate in some systematic and automatic way might be of help in this, although some care was actually given to prevent the acceptance rate getting much under 10 % or over 90 % during annealing. Connected to the transition probability is the temperature setting, as smaller transitions also generally mean smaller changes in energy and smaller required temperature for some acceptance rate. The starting temperature and the shape of the temperature lowering function should in any case probably be optimized in some way for the best results. The stochastic nature of the algorithm can also be a problem, since the resulting energy is not necessarily exactly right even if the conformation is essentially correct. This could be countered by running a local optimization routine at the end of the of the annealing.

References

[1] K. A. Dill and J. L. MacCallum, "The protein-folding problem, 50 years on," *Science*, vol. 338, no. 6110, pp. 1042–1046, 2012.

[2] C. M. Dobson, "Protein folding and misfolding," *Nature*, vol. 426, no. 6968, p. 884, 2003.

[3] F. H. Stillinger, T. Head-Gordon, and C. L. Hirshfeld, "Toy model for protein folding," *Physical review E*, vol. 48, no. 2, p. 1469, 1993.

[4] S. Govindarajan and R. A. Goldstein, "On the thermodynamic hypothesis of protein folding," *Proceedings of the National Academy of Sciences*, vol. 95, no. 10, pp. 5545–5549, 1998.

[5] D. J. Wales and J. P. Doye, "Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110

atoms," *The Journal of Physical Chemistry A*, vol. 101, no. 28, pp. 5111–5116, 1997.

[6] B. Olson, I. Hashmi, K. Molloy, and A. Shehu, "Basin hopping as a general and versatile optimization framework for the characterization of biological macromolecules," *Advances in Artificial Intelligence*, vol. 2012, p. 3, 2012.

6 Appendices

	BBBBB	ABBBA	BABAB	AABAA	BBBBA
Optimal energy	-0,59	-0,40	-0,65	-3,11	-0,067
Annealing reliability	0,55	0,1	1,0	0,6	0,1
Basinhopping reliability	0,1	0,0	0,5	0,65	0,2
	BBBAB	BBBAA	ABBAB	ABBAA	BABAA
	-0,18	0,040	0,026	-0,93	-1,35
	0,25	1,0	1,0	0,3	0,35
	0,0	1,0	1,0	0,1	0,5
	BBABB	ABABA	BAAAB	AAAAA	BBABA
	-0,24	-2,22	-0,52	-4,35	-0,62
	0,9	0,8	1,0	0,35	0,9
	0,15	0,15	0,45	0,45	0,6
	BBAAB	BBAAA	ABAAB	ABAAA	BAAAA
	0,096	-0,55	-1,38	-2,99	-2,20
	1,0	0,95	1,0	0,8	0,95
	1,0	0,55	0,45	0,5	0,55

Table 3: The energy of the optimal configuration from the 20 annealing and basinhopping runs for chains of length 5 and the proportions of runs that reached the correct energy within a deviation of 0,005.

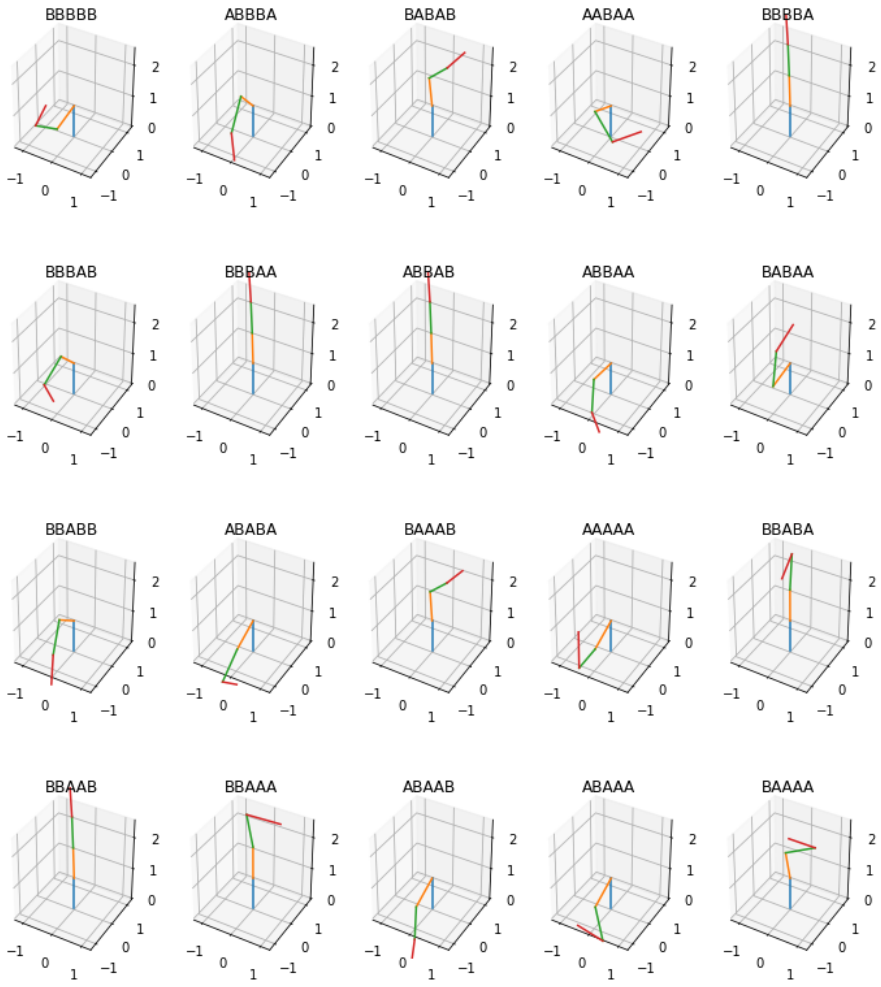

Figure 4: The optimal conformations of the six different chains of length 5 found using simulated annealing. The first acid in the chain always starts at (0,0,0).

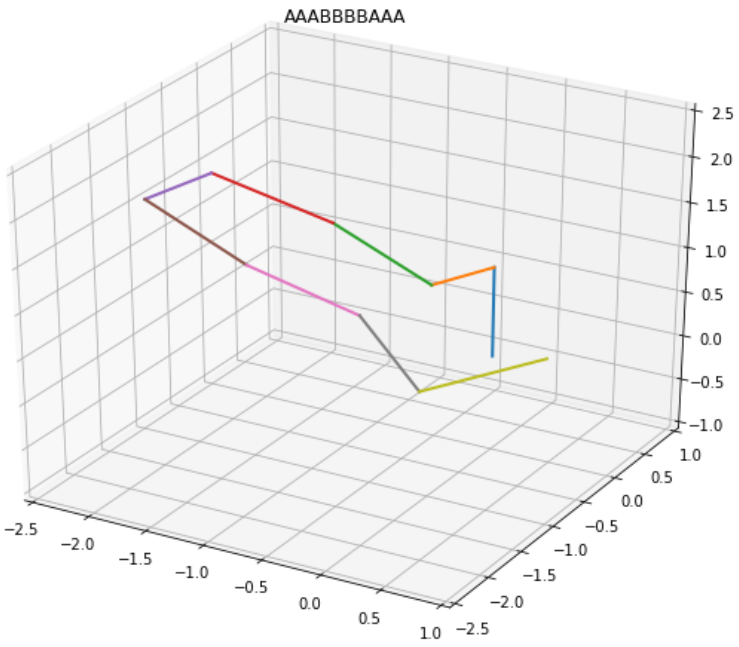

Figure 5: The optimal conformation of the chain AAABBBBAAA found using simulated annealing.

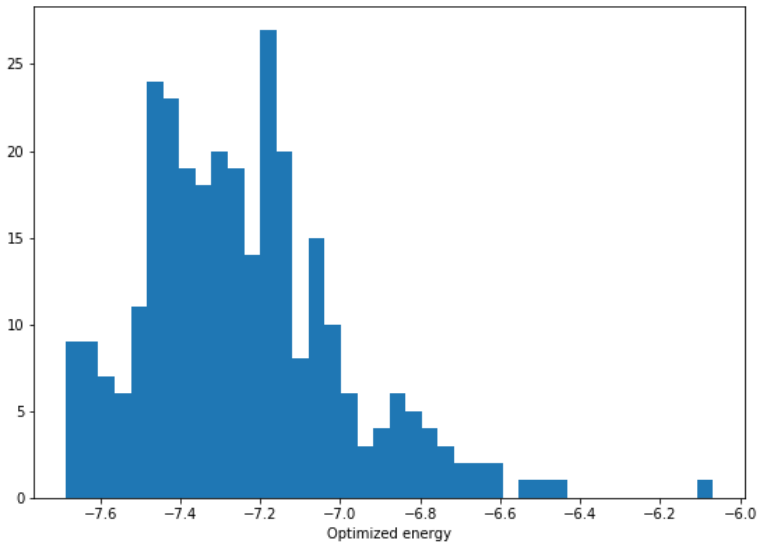

Figure 6: The distribution of the optimized energies got using simulated annealing 300 times on the chain AAABBBBAAA .

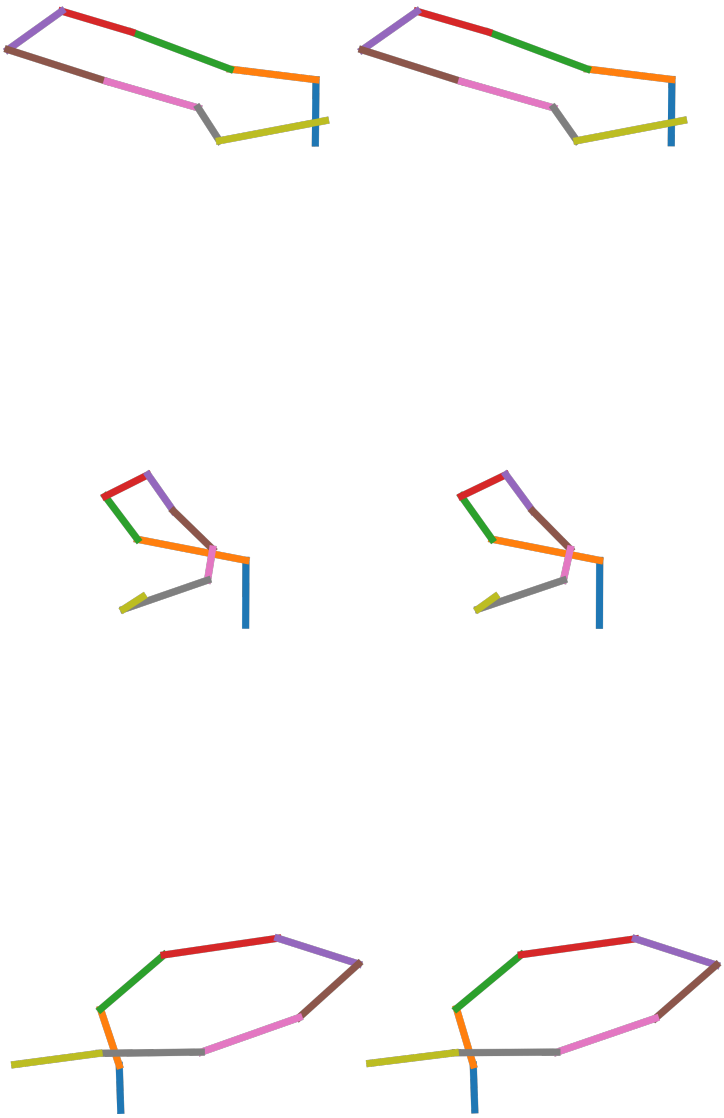

Figure 7: Simple autostereograms for the chain with 10 amino acids. For instructions on how to view autostereograms, see e.g. `http://www.vision-and-eye-health.com/autostereograms.html` (Visited 12.5.2019). The middle one seems to work especially well.