A Go Model of proteins and RNA

by

Amit Garg

Dept of Computer Science and Dept of Biology, Stanford University

Abstract

For large structures Molecular Dynamics is computationally expensive and Monte Carlo is unreliable. Thus, to conduct atomistic simulations we built a Go model, which unites all forces into a single potential. Using the Go model we simulated the folding, unfolding and refolding of proteins and RNA. Our results indicate that a Go model is a valid technique for that purpose. This paper explains how we applied the Go method, showing why it is a feasible alternative to traditional methods and how it can improved.

This work is based on unpublished data from A Garg, B Nakatani, E Sorin in the Pande Lab.

Overview

Understanding how molecules fold and unfold is crucial to understand their function. Modeling the different forces involved in these events can also shed a light on why molecules misfold and cause disease. Much work has been done in protein folding but RNA folding remains largely unexplored. Part of the reason is that RNAs are typically large. Moreover, self-catalytic RNAs (ribozymes) were only discovered a decade ago. Learning more about RNA folding and its catalytic properties could shed new light on a range of domains from evolutionary biochemistry to nanomolecular design.

The golden standard for studying folding is molecular dynamics (MD). MD means calculating the energy and forces and then applying classical mechanics:

$$F_i = -dE_i / dr_i = ma_i \text{ (integrate)}$$

MD has been invaluable in simulating a molecule's trajectory in an energy landscape. It is a powerful technique to model natural processes (eg protein folding) and allows easy comparison with experimental values.

Perhaps the most widely used alternative is Monte Carlo. Monte Carlo is an analytical technique in which a large number of simulations are run using random quantities for uncertain variables. By analyzing the distribution of results one infers which values are most likely.

Both MD and Monte Carlo have their limitations though. In MD, computational time increases exponentially with the number of forces being considered. Thus, the complete folding of even fairly small molecules like a 700-atom RNA hairpin could take thousands of years. Meanwhile, in Monte Carlo the results' reliability is a function of statistical sampling, but the more thorough one is the more computational time it will take. To overcome these barriers scientists either simplify the molecule (eg coarse-grained model) or the technique (eg less precise integration). Another alternative is the Go Model[1].

In a Go model the molecule is placed in a three-dimensional cube. Each atom is its own atomtype and all forces are expressed as a single force potential. By design, the native state is considered the energetic minimum of the model. Because of these simplifications we increase simulation speed but the data has smaller confidence intervals. The original Go model was applied to proteins and observed rapid formation of globular centers, especially if the input is a partially folded state. We believe the model can also be applied to nucleic acids. A Go model of the same 700-atom RNA hairpin can be done in a matter of minutes.

The results of a Go simulation are to be interpreted with caution though. The final state should match the experimental value. Note though the actual trajectory in a Go model might be one of infinite number of possibilities and not correlate with experimentally characterized transition states. The benefit though is the simulation is still atomistic, which can lend powerful insights such as how the folding landscape can be pruned and how different forces interact in shaping secondary structure.

Method

Our potential is akin to the one used by Ueda et al in their original Go model for proteins. To build the model we first obtained structural information for the desired molecule (a pdb file). Parsing the pdb we obtained the interatomic connectivities and distances and imposed the following potential functions:

Table 1: Potential for the Go Model

Bonds (stretching)	$E_{bonds} = \_ k_b (r - r_o)^2$	i to i+1 interactions
Angles (bending)	$E_{angles} = \_ k_a (\_ - \__o)^2$	i to i+2 interactions
Dihedrals (torsion)	$E_{dihedrals} = \_ k_d [1 + \cos(n \_ - \_')]$	i to i+3 interactions
Pair	same as dihedrals	modify i to i+3 interactions except those involving H
Coulomb	$E_{elec} = \__{i \neq j} q_i q_j / r_{ij}$	electrostatic energy (set to zero)
Non-bonded (NB)	$E_{nb} = \__{i \neq j} -A_{ij} / r_{ij}^6 + B_{ij} / r_{ij}^{12}$	non-bonded interatomic potential
Other terms	E_{solv}, $E_{external}$ etc.	unavailable in PDB and not considered

Explanation on Terms:

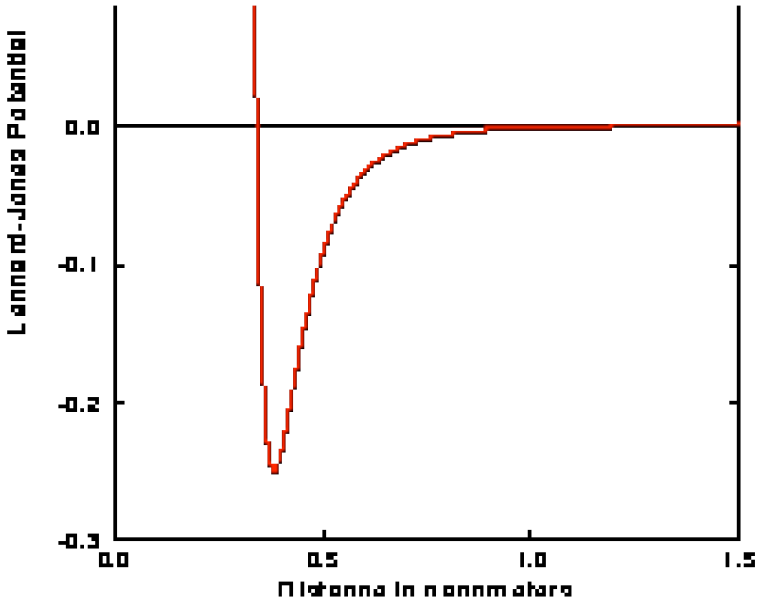
r_o, $\__o$, $\_'$ – distance, angle, dihedral from native structure
k_b, k_a, k_d – bond, angle, dihedral constant f/ force field: oplsaa (proteins), amberN (RNA)
A_{ij}, B_{ij} – calculated from _ and _ (explained below)
Pair – we increase the strength of these interactions to bias our model towards folding
(note: having H interactions also would be too constraining so they are not considered)
Coulomb – still allow for a non-zero potential because we unfold using charges
Other Terms – not considered but we might modify the model in the future

The non-bonded interatomic potential is central to the Go potential. It represents the energetic well for the secondary structure and is modeled as a Lennard-Jones (LJ) function. We defined NB atoms to be those at least 3 residues apart and the total force to be zero at equilibrium. The actual LJ potential can be equated as:

$$\phi(r) = \epsilon \left(\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^{6} \right)$$

where _ = radius at first crossing of zero energy line
_ = constant

Figure 1: the Lennard-Jones Curve



We obtained _ by multiplying r_o from the structure file by $2^{1/6}$ (see the Gromacs manual to understand the conversion) and _ from the force fields[2]. To distinguish strong non-bonded interactions we further adopted the following criteria:

Native interactions – a within 5.5 Ang, multiply _ by 10
Non-native interactions – more than 5.5 Ang

For our simulations we used Gromacs, a highly optimized MD package. We further optimized the code to run a Go simulation. Normally Gromacs calculates NB interactions and places them in atomtype X atomtype matrix, which in our case means a huge n^2 module. Since we subdivided the NB into native and non-native we had it calculate only the non-native and inputted n native interactions. Furthermore, we restricted the non-native interactions – only atoms within a specified cut-off distance will experience a NB force. This effectively reduced the space and memory requirement from $O(n^2)$ to $O(n)$. We also assumed no periodic boundary conditions ie only one molecule so no intermolecular interactions.

Within the framework of our assumptions we tried to make our model similar to reality. The simulations were conducted in water-like viscosity (90 - 100 ps^{-1}) by specifying a tau_t of 0.01:

_ = 1 / _ where _ = frictional coefficient (viscosity)

STP (standard temperature and pressure) conditions were set at 300 K* and 1 atm. For the integrator we used a stochastic instead of deterministic integrations. Also, we updated the neighbor list frequently (every 10 steps) so we could account for the change in position of the atoms. The temperature variables were used to calibrate the model to the molecules behavior in nature (ongoing work).

* Technically "standard temperature" is 298 K but we rounded it to 300 K for convenience.

We also gave random velocities make our model more stochastic. Before every stochastic run we energy minimized the structure, expecting the molecule to settle into the same initial state. Thus, given the energetic well is the same, molecules with different random velocities should still fold to the same native state. We did remove linear rotation to prevent the molecule from moving out of the box; this does not affect the actual folding though. Major global conditions are outlined in Table 2.

Table 2: Some Global Conditions used in Go Model Simulations

Integrator	sd	A leap-frog stochastic dynamics integrator.
dt	0.001	Time step for integration (in ps).
nsteps	10000	Maximum number of steps to integrate.
nstlist	10	Frequency to update the neighbor list (in this case every 10 steps).
ns_type	simple	Check every atom in the box when constructing a new neighbor list every nstlist steps.
coulombtype	cut-off	Twin range cut-off's with rlist <= rvdw <= rcoulomb.
vdwtype	cut-off	Twin range cut-off's with rvdw >= rlist.
rlist	5	Cut-off distance for the short-range neighbor list.
rvdw	5	Distance for the LJ or Buckingham cut-off.
rcoulomb	5	Distance for the Coulomb cut-off.
epsilon_r	1	Dielectric constant.
pcoupl	no	No pressure coupling. This means a fixed box size.
gen_vel	yes	Generate velocities according to a Maxwell distribution.
gen_temp	300	Temperature for Maxwell distribution.
gen_seed	-1	Random generator for random velocities.
comm-mode	linear	Remove center of mass translation.
constraints	hbonds	Only constrain the bonds with H-atoms.
tcoupl	berendsen	Temperature coupling with a Berendsen-thermostat.
tau_t	0.01	Time constant for coupling (viscosity = 10 _ water)
tc_grps	system	Groups to couple separately to temperature bath.
ref_t	300	Reference temperature for coupling.

A Go model is a simplified model. It amalgamates several different potentials – hydrogen bonds, charges etc – into a single potential. Plotkin has shown this approach works well for small protein chains and even for larger molecules with over 1000 atoms[3]. Our innovation was to apply it also to RNA, which are typically much bigger. Comparing our preliminary results to experimentally-determined transition states suggests that the Go model can be applied to nucleics too[4].

Results

We define the following terms:

Stability (100 ps) – Simulations from native state _ native state

Unfolding (10 ps)– simulations from native state _ unfolded state using charges. Charges are an artificial variable and were used primarily to unfold the molecule prior to refolding simulations.

Refolding (100 ps) – Simulations from unfolded state _ refolded state.

Generally we observed that higher constants (k_b, k_a, k_d) caused molecules to fold quicker. However, if these were too high in value the structure became too constrained and caused the simulation to crash. Nonetheless, the driving force in refolding is the LJ interactions. These are considerably weaker than bonds or angles but their sheer number creates a large bulk force. Alternate topologies with weak NB forces not only folded slower but also stabilized at a non-native state, especially in the case of RNA.

Global simulation conditions were sensitive to parameterization. Higher temperatures and higher viscosity increased the fluidity of the model and thus kinetics but not thermodynamics. Moderately high values increased folding but too high actually caused unfolding. nstlist = 10 was also crucial; indeed our early model had an appropriate parameterization but was still unable to refold because the neighbor list was updated too infrequently.

Specifically, we used rmsd as our metric and considered rmsd < 1.0 Ang as indicating strong structural similarity. Table 3 below summarizes the results for four molecules:

Proteins: BBA5, villin headpiece (1qqv.pdb)
RNA: P5Abc region of group I ribozyme (1eor.pdb),
yeast transfer ribonucleic acid (6tna.pdb)

Table 3: Simulation Results for Selected Molecules

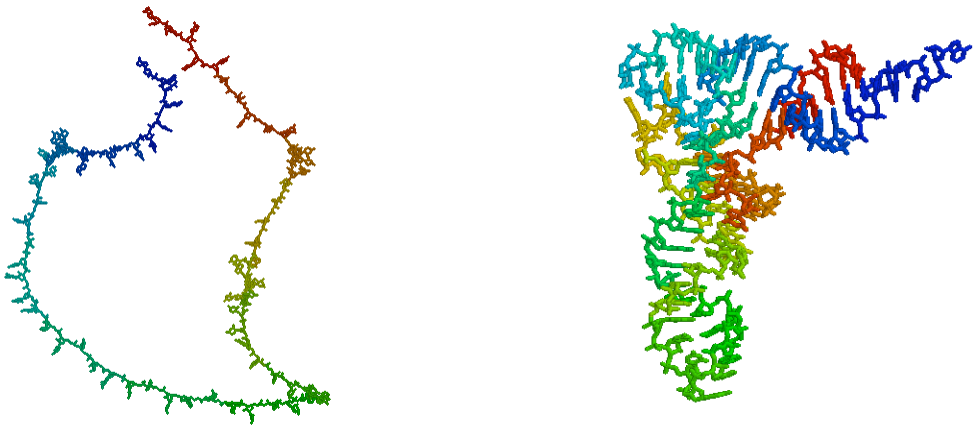
Molecule	# Atoms	Stability (Ang)	Unfolding (Ang)	Refolding (Ang)
BBA5	385	0.30	7.56	0.89
Villin	596	0.30	8.90	0.66
RNA hairpin	718	0.90	16.00	2.66

tRNAphe – 1613 atoms
Our original Go model was folding tRNAphe without the secondary helix formation. Thus, we reduced the bias towards folding by calibrating the native bond _ to temperature, instead of just multiplying it by 10. The folding proceeded slower but gave better results. Here are two refolding simulations in movie format: http://www.stanford.edu/~bjn/6tna/testruns.html.

One can notice the core collapses quickly into an intermediate state. From there, tertiary contacts form within a highly restricted subset of conformational space. This is akin to globular intermediates in protein folding and agrees with theory[5]. Indeed, Russell et al have shown that a fundamental property of folding is compaction from a highly flexible and dynamic set of unfolded conformations to a tightly packed functional structure[6]. Furthermore our simulations passed though transition states similar to those obtained

experimentally, which suggests that different folding trajectories might share subspaces in the energy landscape[7].

Unfolded **Refolded**



Discussion

With the current computational resources traditional simulations of molecules larger than 1000 atoms with traditional MD is unfeasible and with Monte Carlo is unreliable. The Go model provides a simplified but interesting alternative for studying folding trajectories and restricting the folding sample space. Molecules in nature have charges and are constantly interacting with other molecules and ions. Incorporating electrostatic potential into the Go potential and removing intermolecular forces is indeed a major assumption. As Ueda et al found in the original Go paper, the model "simulates the behavior of proteins in essence, but not in fine details."

Another limitation of our Go model is our parameters are too strong. In nature a molecule randomly crosses the energy barrier and folds or unfolds in the scale of milliseconds to minutes[8]. In our model, we folded within a few nanoseconds and could not detect unfolding in any of the runs. Given that we fold so quickly most likely we will not unfold even in the millisecond to minute scale. Indeed, our model biased is biased towards folding because a simulation of a minute would take a few months at best. We have also realized that constraining the backbone bonds, angles and dihedrals caused the molecules to fold even faster and might use this to speed simulations for larger molecules.

To keep the benefits of the bias without distorting accuracy we are working to calibrate the model to temperature. This allows our simulation to behave more similar to nature. Indeed, preliminary results indicate that our original parameterization of tRNAphe was constraining the molecule to a natural temperature of 100 K. By calibrating to room temperature we might be able to observe some unfolding.

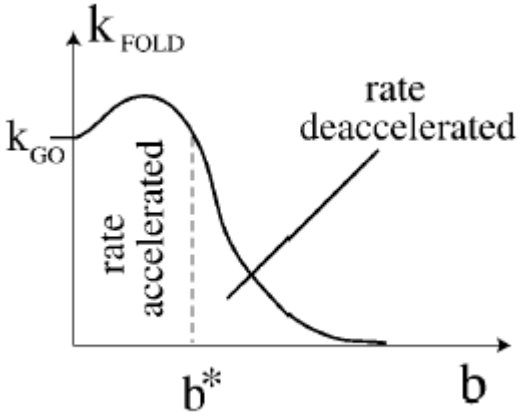
One key feature of Go simulation is that it does not map to real time trajectories should be interpreted with caution. Moreover, the Go potential integrates over discrete units whereas in nature folding occurs as a continuous function[9, 10]. In our model, a time step of 1 fs was usually fine but 2 fs typically caused simulations to crash because the atoms came too close together. Ideally we want to integrate with a smaller time step and for longer time to approach reality.

Another limitation is based on the software package itself. Gromacs is powerful but not easy to use or modify. Onuchic has shown that folding depends largely on topology[11, 12]. Given the number of factors to consider For instance, hydrogen bonding and hydrophilic interactions are crucial to RNA secondary structure but Gromacs does not distinguish these from other factors in the Go potential. Currently only oplsaa is built into Gromacs; Sorin has been working to incorporate amberN to help in this issue.

Regarding thermodynamics we observed a definite bias towards the native state. Kinetics though is a more contentious issue. Some claim that optimal folding takes place in an entropic dominated time scale. Thus, within a certain temperature range proteins will fold fast roughly independent of size and structural spoecificites[13]. Others claim that folding depends largely on geometry and not really dependent on size[14]. These two predictions are not mutually exclusive though and we do not have enough results yet to confidently support any of them.

NB forces provide an interesting consideration for Go kinetics. NB forces are both repulsive and attractive and assist or hinder folding. Figure 2 illustrates the latter two regimens:

Figure 2: NB forces assisting or hindering folding



b = number of NB interactions
b* = turning point.

adapted from Plotkin

This diagram shows that a larger number of NB forces does not necessarily mean a faster or slower folding rate. For instance, too many attractions can actually oppose folding if they cancel each other. In the tRNA NB forces working in unison are probably responsible fro the rapid collapse of the core.

Conclusion

Atomistic MD simulations are ideal but currently impractical for large structures. Alternatives like Monte Carlo provide a statistical sampling but are not nearly as universal. Traditional simplifications like coarse-grained simulations typically only highlight general patterns. In this context, the Go model is a feasible alternative to simulate the folding and refolding of proteins and RNA.

Go results must be interpreted with discretion however. A folding trajectory might lead to a local minimum and even if it leads to the natural state, the trajectory is not necessarily the one molecules follow in real life. There is growing evidence that molecules following several different folding trajectories though and the Go model might be useful to tease the folding space. Finally the Go model is useful to understand how different forces interact to produce secondary structure, and might be used for structure prediction in the future.

We are currently refining the model and hoping to try it soon on larger ribozymes. Our ultimate goal is to perform an all-atom simulation of the Tetrahymena ribozyme, which contains approximately 11,000 atoms. No group has yet simulated such a large molecule atomistically. Furthermore, Nakatani developed a 36-sphere coarse-grain model of the ribozyme in the past and Chu has recently published experimental data on it[15]. Thus, we will hopefully have a good standard against which to judge our future work.

Acknowledgments

Thank you to Eric Sorin and Brad Nakatani for their helpful comments and suggestions.

Links

Protein Data Bank (PDB) -- http://www.rcsb.org/pdb/
Tinker -- http://dasher.wustl.edu/tinker/
Gromacs -- http://www.gromacs.org/

References

1) Y Ueda, H Taketomi, N Go. *Studies on Protein Folding, Unfolding, and Fluctuations by Computer Simulation. II. A Three-Dimensional Lattice Model of Lysozyme*. Biopolymers vol 17, 1531-1548 (1978).

2) D van der Spoel, A Buuren, E Apol, Pieter Meulenhoff, D Tieleman, A Sijbers, B Hess, K Feenstra, E Lindahl, R Drunen, H Berendsen. *Gromacs manual version 3.1.1*. ftp://ftp.gromacs.org/pub/manual/3.1/manual-a4-3.1.1.pdf (2002)**.**

3) S Plotkin. *Speeding Protein Folding beyond the Go Model: How a Little Frustration Sometimes Helps*. Proteins: Structure, Function, and Genetics 45, 337-345 (2001).

4) X Zhuang L Bartley, H Babcock, R Rusell, T Ha, D Herschlag, S Chu, *Single Molecule Study of RNA Catalysis and Folding*. Science 288, 2048-2051 (2000).

5) Z Cai, I Tinoco, Jr. *Solution Structure of Loop A from the Hairpin Ribozyme from Tobacco Ringspot Virus Satellite*. Biochemistry 35, 6026-6036 (1996).

6) R Russell I Millett, M Tate, L Kwok, B Nakatani, S Gruner, S Mochriw, V Pande, S Doniach, D Herschlag, L Pollack. *Rapid compaction during RNA folding*. PNAS vol 99, 4266-4271 (Apr 2, 2002).

7) P Rubert, A Massey, S Sigurdsson, A Ferré-D'Amaré. *Transition State Stabilization by a Catalytic RNA*. Science vol 298, 1421-1424 (Nov 15, 2002).

8) S Butcher, F Allain, J Feigon. *Solution structure of the loop B domain from the hairpin ribozyme*. Nature Structural Biology vol 6, 212-216 (Mar 1999).

9) M Cheung, A Garcia, J Onuchic. *Protein folding mediated by solvation: Water expulsion and formation of the hydrophobic core occur after the structural collapse*. PNAS vol 99, 685-690 (Jan 22, 2002).

10) M Fedor. *Structure and Function of the Hairpin Ribozyme*. JMB 297, 269-291 (2000).

11) C Clementi, P Jennings, J Onuchic. *How native-state topology affects the folding of dihydrofolate reductase and interleukin-1_*. PNAS vol 97, 5871-5876 (May 23, 2000).

12) C Clementi, H Nymeyer, J Onuchic. *Topological and Energetic Factors: What Determines the Structure Details of the Transition State Ensemble and "En-route" Intermediates for Protein Folding? An Investigation for Small Globular Proteins*. JMB vol 298, 937-953 (2000).

13) A Gutin, V Abkevich, E Shahknovich. *Chain-Length Scaling of Protein Folding Time*. Phys Rev Letter vol 77, 5433 (1996).

14) H Jang, C Hall, Y Zhou. *Protein Folding Pathways and Kinetics: Molecular Dynamics Simulations of _-Strand Motifs*. Biophysical Journal vol 83, 819-835 (August 2002).

15) R Russell, X Zhuang, H Babcock, I Millett, S Doniach, S Chu, D Herschlag. *Exploring the folding Landscape of a Structured RNA*. Proc Natl Acad Sci USA vol 99, 155-60 (2002).