All-Atom CSAW: An Ab Initio Protein Folding Method

Weitao Sun[1*]

[1] Zhou Pei-Yuan Center for Applied Mathematics, Tsinghua University, Beijing, 100084, China
* Corresponding author: sunwt@tsinghua.edu.cn

Abstract Conditioned Self-Avoiding Walk (CSAW) was first developed as a tool to simulate protein folding Based on CSAW algorithm, All-atom Conditioned Self-Avoiding Walk (AA-CSAW) was developed around 2007. The polypeptide chain is simulated as effectively rigid cranks -Ca-CO-NH- units lined by covalent bonds. Bond lengths and bond angles are set as fixed optimal values. All-atom amino acid sidechain is attached to very Ca atom. The structure of polypeptide is fully described by backbone dihedral angles φ, ψ and the sidechain dihedral angles χ. A trial structure is randomly generated by pivoting the polypeptide chain and sidechains. In the pivot algorithm, the backbone dihedral angles φ, ψ for each residue are chosen according to probability distributions in Ramachandran plot. The dihedral angle distributions are improved by 3-residue fragment set investigation. The effective energy of protein structure is constructed by considering hydrophobic effect, desolvation effect and hydrogen bonding interaction. An appropriate three dimensional structure is accepted with a probability according to Metropolis scheme. In order to evaluate the accepted structures in Monte Carlo simulations, the ratio of secondary structure content to radius of gyration is introduced. CASP09 target example shows that AA-CSAW is an efficient and promising ab initio method.

Keywords protein folding simulation, self-avoiding walk, coarse-grained, sidechain atom

1. Introduction

In physiological conditions, globular protein folds from randomly coiled polypeptide chain into a characteristic three-dimensional structure in water solution. Protein folding is a stochastic process and there are huge amount of protein molecules in organism. The observable macroscopic properties have microscopic interpretations based on collective molecule behaviors. Meanwhile, individual protein molecule undergoes a Brownian motion and it's hard to describe the movements of each molecule. In addition, protein structure is at the edge of thermo-equilibrium. Delicate balance between entropy-enthalpy exist throughout the folding process. We believe that statistical thermodynamics is a prior way for protein folding problem.

Polypeptide chains in solutions incessantly change shape and position by thermal agitation. This Brownian motion can be characterized in more quantitative fashion by the use of phenomenological models. One such model is summarized by the Langevin equation of motion[1, 2]. Monte Carlo (MC) simulation samples typical configurations from a Boltzmann distribution determined by potential energy and temperature of the system. Monte Carlo method solves the stochastic models without consideration of the analytical representations of the system. It is clear that the Monte Carlo model does not represent the dynamic behavior of the real system directly. According to the ergodicity theorem, "time average" will converge to the "configuration average" in a large system. Estimations from MC simulations often correspond very well with those from MD simulations.

In All-tom CSAW method[3]., we first set up an initial unfolded structure. Then the structure is pivoted by randomly choosing backbone dihedral angles ϕ,ψ and sidechain torsion angle χ. This candidate structure is checked by self-avoiding walk criteria to make sure that there are no atom overlaps. The structure energy is calculated for the pivoted peptide chain. A Metropolis scheme is used to determine if the new structure should be accepted. If accepted, this pivoted structure is saved and used as the starting point for the next loop. Otherwise, the structure is restored to the one before pivot. Then a new pivot is carried out in the next loop.

2. “Crank” model of Amino acid

Amino acids are molecules containing an amine group, a carboxylic acid group and a side chain that varies between different amino acids. The amine and carboxylic acid groups of amino acids react to form amide bonds. One amino acid molecule can react with another and become joined through an amide linkage. This polymerization of amino acids yields the newly formed peptide bond and a molecule of water.

The protein is a linear polymer of the 20 different kinds of amino acid, which are linked by peptide bonds. All of the 20 amino acids have in common a central carbon atom (C_α) to which are attached a hydrogen atom, an amino group, and a carboxyl group (COOH). What distinguishes one amino acid from another is the side chain attached to the C_α. The main-chain atoms are a carbon atom C_α, an NH group bound to C_α, and a carbonyl group C’=O, where the carbon atom C’ is attached to C_α. The basic repeating unit along the main chain is thus (NH- C_α-C’=O), which is the residue of the common parts of amino acids after peptide bonds have been formed.

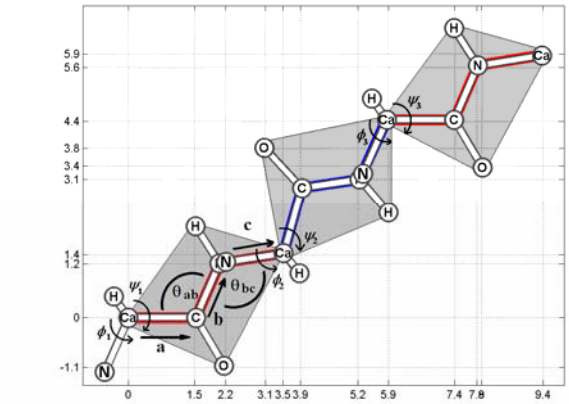

Figure 1. Crank chain model of protein backbone structure

The peptide bond tend to be planar and the rigid peptide dihedral angle (the bond between C’ and N) is always close to 180 degrees (Figure 1). Here C’ indicate the carbon atom bonded to C_α. Because of the double-bond unrotatable feature of peptide bond, the atoms C_α-C’=O-N are constrained in a rigid plane, called peptide plane. There are obvious repeating patterns -Cα-C'=N-Cα in protein backbone. The interlink within these atoms form a crank-shaped rigid body. Thus, we introduce a ‘crank’ unit –Cα-C1=N- in CSAW. Crank is a simplification of amino acid residues. The Cα atom is located at origin point.

3. Add sidechain atoms to residue

What distinguish protein from common polymers are the diversity of residue sidechains. There are 20 natural amino acid residues in protein. They can be classified into several groups, such as hydrophobic, charged, polar and so on. Some works show that hydrophobic/hydrophilic residue types are enough to create secondary structures (such as alpha helix). However, the importance of other subtle properties of amino acid sidechains are still underestimated. The delicate sidechain differences may shed light to the folding of native structure.

As for coarse-grained model, all atom sidechains are substantially important. The single sphere sidechain model is good. But sometimes spurious overlaps happen when anisotropic residue sidechains come close to each other. In single ball sidechain model, many important dense configurations will be ignored, which prevent protein structure from search correct folding path in conformation space. Protein folding simulations of real proteins make demand of an all-atom

sidechain CSAW method. We developed the all-atom self-avoiding walk method as a coarse-grained ab initio protein folding simulation method with atom details[3].
Since crank model can provide atom locations for backbone atoms, the central problem is how to determine the sidechain atom coordinates if the atom coordinates are known for a backbone structure in arbitrary orientation. Thanks for the knowledge of amino acid structure, we have the atom coordinates for sidechain in some special orientation. As a consequence, we can determine the sidechain atom coordinates by matching amino acid to the backbone of a crank.
As the structure of 20 amino acids are well determined by experiment observation, we have the atom coordinates for any type of residues, including backbone $\mathbf{X}_{BB}^{obs}$ and sidechain $\mathbf{X}_{SC}^{obs}$. The only problem is that the observed amino acid structure are usually not in the same orientation as in crank model. If the backbone parts N-C_α-C-O of observed amino acid structure overlap with crank model, it is obvious that the crank sidechain atom will be determined by $\mathbf{X}_{SC}^{crank} = \mathbf{X}_{SC}^{obs}$.

4. Pivot algorithm

The protein chain is simulated by a series of cranks connected at C_α atom. The chain conformation will change when the torsion angle φ and ψ are modified. In order to create a new conformation in AA-CSAW, we pivot the chain at certain randomly selected C_α atom by changing φ and ψ to specified values (Figure 2). The crank chain pivot procedure can be summarized as following steps.

1. Start with initial chain.
2. Choose a crank randomly as pivot point and change the torsion angles
3. Rotate end portion about pivot point.
4. Check atom overlaps. If no overlap, accept and update chain conformation.
5. If overlap, go to **2.**

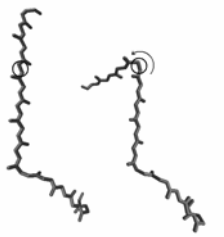

Figure 2. Pivot polypeptide chain by changing torsion angles

5. Non-covalent Interactions and structure energy

In AA-CSAW, we consider the most important non-covalent interactions in water solution, such as hydrophobic effect, hydrogen bonding.
Hydrophobic residues, such as Alanine, Valine and Phenylalanine, tend to congregate in an aqueous (i.e. water) environment. The absence of hydrogen bonding between water and non-polar groups constitutes an important source of the protein stability in aqueous solution, known as hydrophobic effect. With the help of such effect, a hydrophobic core is usually formed in globular protein. The assembly of hydrophobic residues is thus considered as the driving force of the collapse of peptide chain[4].
Hydrophobicity is usually expressed as the Gibbs energies of transfer from water into the reference state. When globular protein fold from denature to native structure, hydrophobic residues come together and free energy decrease. The transfer hydrophobic amino acid residues from cyclohexane

into water is energetically costly. Thus the burial of hydrophobic sidechains in the folding reaction is energetically favorable.

In order to model this folding process, energy decrease caused by hydrophobic residue aggregation is simulated by residue contact degree. The hydrophobicity of amino acids are represented by h. When h is positive, the residue is hydrophobic. Otherwise, it is hydrophilic. Table 1 shows the hydropobic properties of 20 amino acids. If residue m has multiple contact neighbors, the hydrophobic energy variation is the summation of residue hydrophobicities.

Table 1. Hydrophobicity of 20 amino acids

Amino acid	3-letter codes	1-letter codes	Hydrophobicity	Amino acid	3-letter codes	1-letter codes	Hydrophobicity
Alanine	Ala	A	0.05	Leucine	Leu	L	0.15
Arginine	Arg	R	-0.42	Lysine	Lys	K	-0.21
Asparagine	Asn	N	-0.28	Methionine	Met	M	0.04
Aspartic Acid	Asp	D	-0.35	Phenylalanine	Phe	F	0.06
Cysteine	Cys	C	0.01	Proline	Pro	P	
Glutamic Acid	Glu	E	-0.25	Serine	Ser	S	-0.21
Glutamine	Gln	Q	-0.21	Threonine	Thr	T	-0.15
Glycine	Gly	G	0.00	Tryptophan	Trp	W	0.03
Histidine	His	H	-0.17	Tyrosine	Tyr	Y	-0.03
Isoleucine	Ile	I	0.15	Valine	Val	V	0.13

Hydrophilicity at 25℃ is relative to glycine, and is based on the partitioning of a sidechain analogue between the two states [5-8].

Hydrogen bond (H-bond) contains both positive (H-donor) and negative (H-acceptor) partial charges. It represents a combination of covalent and electrostatic interactions, but the main component is the electrostatic attraction between hydrogen donor and acceptor. Hydrogen bonding interactions between backbone O atom and N-H atoms has different stabilities in different situations. The Gibbs energy contributions per hydrogen bond in the interior of proteins are estimated to be 10-60 kJ/mol[5, 9, 10]. While the hydrogen bond interaction in protein structure is much weaker at the surface, where the relative permittivity of water is close to 80.

6. Example of CASP09 target simulation

Here we show the CASP09 target T619 (3NRW) as an folding simulation example. T619 has 111 amino acids and is composed of five helices (Figure 3(A)). The initial structure is unfolded long chain. The final predicted 3-D structure by CSAW is shown in Figure 3(B).

In addition to the final structure, we calculate many other important parameters for the whole folding process. The hydrophobic energy and hydrogen bond energy are shown in Figure 3(A). Although the electrostatic energy is not involved in current AA-CSAW as a driving force, we do calculate it for charged residues through Coulomb's law.

The interesting features of energy curves are (1) Hydrophobic (hp) energy decreases fast within first 300 steps. Then hp curve fluctuates around a stable value. This indicates a fast structure collapse in the very early folding stage. Afterwards the hydrophobic residue clusters do not grow greatly. (2) Hydrogen bond (hb) energy curve descends continuously during the whole folding process. The

decrease trend implies a sustained hydrogen bond formation (see Figure 3(B)), which is an indicator of increasing secondary structure. (3) The cross of hp and hb energy curves show that hydrophobic effect dominate the early folding stage and the hydrogen bonding interaction controls the following stage after a transition state around step 3500.

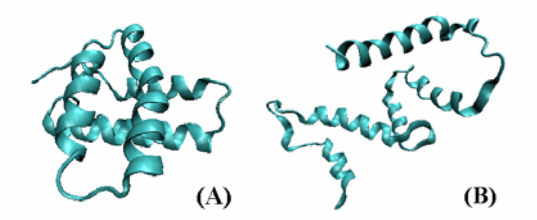

Figure 2. Native and predicted structure of target T619

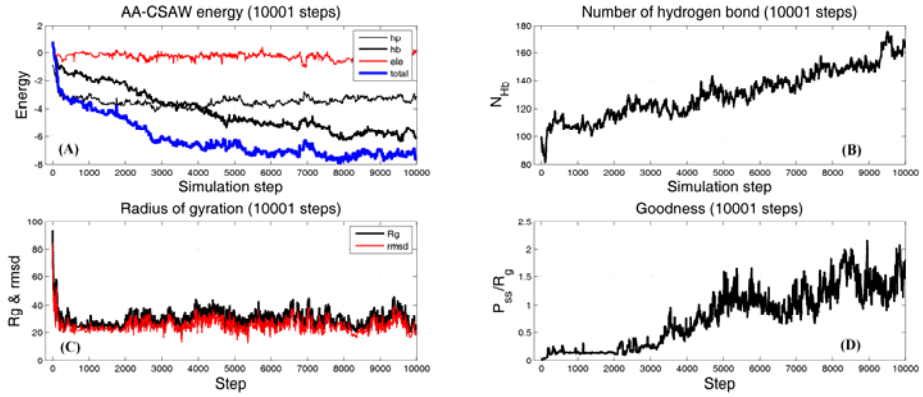

Figure 3. Simulation results for 10001 steps of target T619. (A) Hydrophobic (hp) energy, hydrogen bond (hb) energy, electrostatic (ele) energy and total energy. (B) Number of hydrogen bond. (C) Predicted structure radius of gyration and RMSD. (D)Structure evaluation parameter

The radius of gyration (Rg) and RMSD curves show obvious stages during the folding process. (1) Before update step 300, there is a fast collapse of Rg. This is a reasonable result from fast mainchain collapse driven by hydrophobic effect. In this stage, protein backbone can pivot freely because of relatively less residue repulsive forces. As residues packing to each other, the hydrophobic cores are formed and there are many hydrogen bonds appear. Hydrophobic energy and hydrogen bonding energy have a sharp decrease during the first 300 steps, so as the number of hydrogen bonds in this fast collapse stage (Figure 3(B)). Considering definition of CSAW energy, protein backbone collapse fast to a globular structure containing may hydrophobic cores. This globular shaped structure becomes more stable with the help of hydrogen bonds. (2) From update step 300 to 2000, there is a slow packing process. Protein structure Rg decrease slowly from 30 Å to 20 Å. The decrease of hydrophobic and hydrogen bonding energy also slows down. This means that it become difficult for protein to find a ‘comfortable’ structure with lower energy.

As the hydrophobic residue cluster grow to a certain size, the increase of hydrophobic effect will slow down. This is caused by the dewetting effect near large hydrophobic cluster surface. On the contrary, hydrogen bonding interaction keep increasing as more H-bond appear. Hydrogen bond energy take the place of hydrophobic energy. Since hydrophobic effect cannot hold the globular structure as tight as before, protein has more chance to expand in volume and accommodate more hydrogen bonds. Consequently, the radius of gyrations show large fluctuation after step 2000.

Large Rg fluctuation makes protein structure much more flexible, which provides a better way to search native structure in conformation space. Figure 3(D) shows that the structure become better and better after the turning point around step 2000. Whenever the control of hydrophobic effect relaxes, the structure optimization speeds up. Thus, a modifiable hydrophobic effect, which is the true physical picture in protein folding, is much reasonable than a constant one.

7. Conclusions

The newly developed AA-CSAW method integrates both coarse-grained crank backbone model and rotatable sidechain atoms to improve protein structure prediction. The deep physical understandings of noncovalent interactions lead to incorporation of solvent effect in hydrophobic and hydrogen bond energy calculation. To our best knowledge, this is the first time to propose such improved algorithm in coarse-grained ab initio method. The cluster-size dependent hydrophobic effect makes the structure much more flexible. Thus, the hydrophobic and hydrogen bonding interactions are well balanced as two stages in folding process. CASP09 target example shows that the AA-CSAW method is now ready to simulate some real and large structures.

Acknowledgements

I thank Professor Kerson Huang for interesting and helpful discussions. The work in this paper is sponsored Tsinghua University Initiative Scientific Research Program (20101081751). Part of this work is supported by Institute of Advanced Studies at Nanyang Technological University.

References

[1] K. Huang. PROTEIN FOLDING AS A PHYSICAL STOCHASTIC PROCESS. Biophysical Reviews and Letters, 3 (2008) 1-18.

[2] K. Huang. CONDITIONED SELF-AVOIDING WALK (CSAW): STOCHASTIC APPROACH TO PROTEIN FOLDING. Biophysical Reviews and Letters, 2 (2007) 139-54.

[3] W. Sun. Protein folding simulation by all-atom CSAW method. IEEE International Conference on Bioinformatics and Biomedicine, 2 (2007) 45 - 52.

[4] W. Kauzmann. Some Factors in the Interpretation of Protein Denaturation. Adv Protein Chem, 14 (1959) 1-63.

[5] P. Dauber, Hagler A. T. Crystal packing, hydrogen bonding, and the effect of crystal forces on molecular conformation. Accts Chem Res, 13 (1980) 105-12.

[6] P. L. Privalov, Makhatadze G. I. Contribution of hydration to protein folding thermodynamics. 2. The entropy and Gibbs energy of hydration. J Mol Biol, 232 (1993) 660-79.

[7] A. Radzicka, Wolfenden R. Comparing the polarities of the amino acids: side chain distribution coefficients between the vapor-phase, cyclohexane, 1-octanol, and neutral aqueous solution. Biochemistry 27 (1988) 1664-70.

[8] T. E. Creighton. Proteins-Structures and molecular properties. 2 ed, W. H. Freeman and Company, New York, 1993.

[9] A. T. Hagler, Dauber P., Lifson S. Consistent force field studies of intermolecular forces in hydrogen bonded crystals. III. The C=O...H-O hydrogen bond and the analysis of the energetics and packing of carboxylic acids. J Am Chem Soc, 101 (1979) 5131-41.

[10] G. I. Makhatadze, Privalov P. L. Contribution of hydration to protein folding thermodynamics. 1. The enthalpy of hydration. J Mol Biol, 232 (1993) 639-59.