ISSN: 2347-3487

The Second Law, Gibbs FreeEnergy, Geometry, and Protein Folding

Yi Fang[1]
1Department of Mathematics, Nanchang University,
& Department of Mathematics, The Australian National University
yifang@ncu.edu.cn, yi.fang3@gmail.com, yi.fang@anu.edu.au

The fundamental physical law of protein folding is the second law of thermodynamics. The key to solve protein folding problem is to derive an analytic formula of the Gibbs free energy. It has been overdue for too long. Let U be a monomeric globular protein whose M atoms $(\mathbf{a}_1,\cdots,\mathbf{a}_M)$ are classified into hydrophobicity classes $H_1,\cdots,H_H$, $H \geq 2$. For each conformation $\mathbf{X} = (\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_M)$ of U, we apply quantum statistics to the corresponding single molecule thermodynamic system $\mathsf{T}_{\mathbf{X}}$ to obtain Gibbs free energy $G(\mathbf{X};\mathsf{U},\mathrm{En}_{\mathbf{N}})$ of protein folding in physiological environment $\mathrm{En}_{\mathbf{N}}$,

$$G(\mathbf{X};\mathsf{U},\mathrm{En}_{\mathbf{N}}) = \omega_e V(\Omega_{\mathbf{X}}) + d_w \omega_e A(M_{\mathbf{X}}) + \sum_{i=1}^{H} \omega_i A(M_{\mathbf{X}i}) + \sum_{1\leq A<B\leq M} \frac{q_A q_B}{4\pi\varepsilon_0 |\mathbf{x}_A - \mathbf{x}_B|}$$

where $V(\Omega_{\mathbf{X}})$ is the volume of $\Omega_{\mathbf{X}}$, a region enclosed by the molecular surface $M_{\mathbf{X}}$, d_w the diameter of a water molecule, ω_e and ω_i's chemical potentials per unit volume and area, and $A(M_{\mathbf{X}})$ and $A(M_{\mathbf{X}i})$ the areas of $M_{\mathbf{X}}$ and $M_{\mathbf{X}i} = \{\mathbf{z} \in M_{\mathbf{X}} : \mathrm{dist}(\mathbf{z}, \cup_{\mathbf{a}_j \in H_i} B(\mathbf{x}_j, r_j)) \leq \mathrm{dist}(\mathbf{z}, \cup_{\mathbf{a}_j \notin H_i} B(\mathbf{x}_j, r_j))\}$. The $G(\mathbf{X};\mathsf{U},\mathrm{En}_{\mathbf{N}})$ and its gradient $\nabla G(\mathbf{X};\mathsf{U},\mathrm{En}_{\mathbf{N}})$ not only reduce the protein structure prediction to a pure mathematical problem of finding minimizers of an analytic function $G(\cdot;\mathsf{U},\mathrm{En}_{\mathbf{N}}) : \mathsf{R}^{3M} \to \mathsf{R}$, but also supply new insights in understanding the kinetics of the protein folding process.

Council for Innovative Research
Peer Review Research Publishing System
Journal of Advances in Physics
Vol 3, No.3
editor@cirworld.com
www.cirworld.com, member.cirworld.com

ISSN: 2347-3487

Decades of experiments by many researchers proved that once the peptide chain of a natural protein is put in correct environments it will spontaneously fold to its native structure. Therefore, the guiding fundamental physical law must be the second law of thermodynamics. Anfinsen summarized this as the thermodynamic principle (he modestly called it hypothesis) of protein folding, that the native structure has the minimum Gibbs free energy and only depends on the peptide chain of the protein in physiological environment [1]. Thus, a cross section of complicated life phenomena, the protein folding, is reduced to a physical problem and should and can be solved accordingly.

Theoretically, all problems such as structure prediction and mechanics of folding process will be answered once we know the Gibbs free energy of protein folding, $G(\mathbf{X};\mathsf{U},\mathrm{En}_{\mathbf{N}})$, where $\mathbf{X}=(\mathbf{x}_1,\cdots,\mathbf{x}_M)\in\mathsf{R}^{3M}$ is a conformation of a protein U which has M atoms $(\mathbf{a}_1,\cdots,\mathbf{a}_M)$ and $\mathbf{x}_i\in\mathsf{R}^3$ is the atomic (nuclear) center of $\mathbf{a}_i$. According to Anfinsen, the physiological environment $\mathrm{En}_{\mathbf{N}}$ consists of elements such as "solvent, p H, ionic strength, presence of other components such as metal ions or prosthetic groups, temperature, and other", [1]. We think that pressure belongs to the other. In fact, because constant pressure and variant volume, the second law takes the version of minimum of Gibbs free energy. We will derive $G(\mathbf{X};\mathsf{U},\mathrm{En}_{\mathbf{N}})$ for a monomeric globular protein U via quantum statistical mechanics.

We start with the observable physical quantity, the electron density distribution function [2],

$$p_{\mathbf{X}}(r)=p(r;\mathbf{X})=N\sum(\mathrm{spins})\left\{\int\mathrm{d}\tau_2\int\mathrm{d}\tau_3\cdots\int\mathrm{d}\tau_N\Psi^*(\mathbf{x};X)\Psi(\mathbf{x};X)\right\},$$

where $\Psi(\mathbf{x};\mathbf{X})$ is the wave function of the Born-Oppenheimer approximation to the Hamiltonian of one molecule of U, and N is the number of electrons in U.

Since in natural, nascent peptide chains already have their peptide bonds and covalent bonds in residues formed, we will not discuss the bond lengths and angles. Instead, we assume that the values of those covalent bond lengths and angles in $\mathbf{X}$ are very close to the standard bond lengths and angles.

To apply statistical mechanics, we have to create a thermodynamic system $\mathsf{T}_{\mathbf{X}}$ tailor made for $\mathbf{X}$. $\mathsf{T}_{\mathbf{X}}$ is a region in R^3 that contains $P_{\mathbf{X}}=\cup_{i=1}^{M}B(\mathbf{x}_i,r_i)$ and its immediate environment, leaving everything in $\mathsf{R}^3\backslash\mathsf{T}_{\mathbf{X}}$ as the heat bath. Here r_i s are van der Waals radii (taking as constants) and $B(\mathbf{x}_i,r_i)$ is the ball in R^3 centered at $\mathbf{x}_i$ of radious r_i. All state functions of $\mathsf{T}_{\mathbf{X}}$ will depend on the following: 1. the conformation $\mathbf{X}$ hence $P_{\mathbf{X}}$; 2. the immediate environment ($\mathrm{En}_{\mathbf{N}}$) of $P_{\mathbf{X}}$ inside $\mathsf{T}_{\mathbf{X}}$; and 3. the peptide chain of the protein U. To be realistic, these are general requirments for any attempting of creating a Gibbs free energy functon of protein folding.

Because of $\mathsf{T}_{\mathbf{X}}$ is tailor made for $\mathbf{X}$, requirment 1 is automatically satisfied. For reuqirement 3, only monomeric globular proteins can be assumed that in the immediate environment of $P_{\mathbf{X}}$ there are no other large objects except water molecules, hence here we consider only this class of proteins for the simplicity of environemnet. Note that the method itself is general, only that for complicated environment the derived formula will also more complicated than that obtained here. For the requirment 2, the dependence on the peptide chain of U is via the electronic density distribution function $p_{\mathbf{X}}$ that indicates how the $P_{\mathbf{X}}$ will interact with the immediate environment, in our case, water. For which we need to discretize $p_{\mathbf{X}}$ with general knowledge of amino acids. It was well-known as early as the 1920's that proteins are multi-polar or "bristling with charges" as described in [3], resulting in different atomic groups have different hydrophobicity levels, say, there are $H\geq 2$ hydrophobicity classes $H_1,\cdots,H_H$. We can assign an atom $\mathbf{a}_k$ into one hydrophobic class H_i if $\mathbf{a}_k$ belongs to an H_i atomic group. For example, we may assume that the classification is as in [4] where there are $H=5$ classes, C, O/N, O^-, N^+, S. Unlike in [4], we also classify every hydrogen atom into one of the H hydrophobicity classes. The atomic space distribution of these hydrophobicity classes are highly depending on $\mathbf{X}$ and the peptide chain of U. Exploiting these space distributions gives a way of applying the $p_{\mathbf{X}}$ while not being able to calculate it.

ISSN: 2347-3487

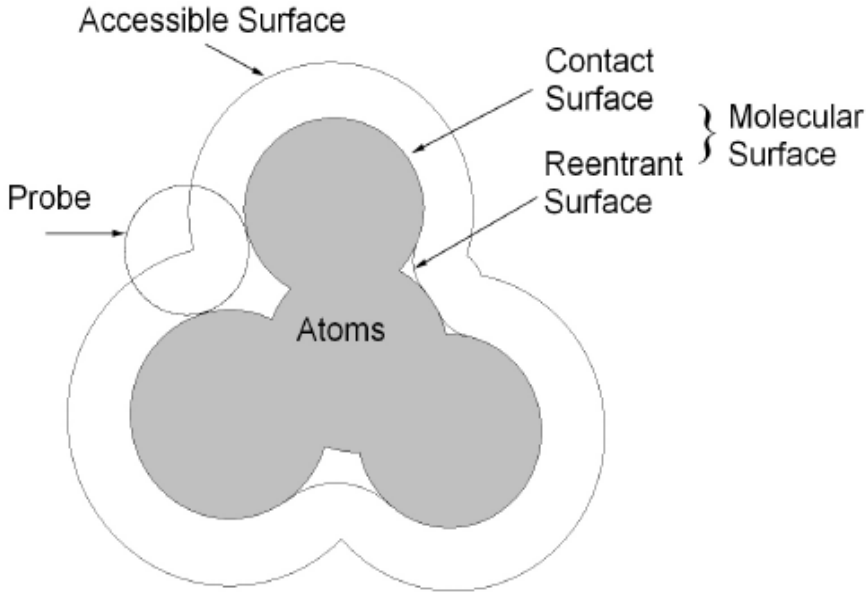

Figure 1: Molecular surface.

To describe the formula, we have to do some preparations. Rolling a probe sphere of radius r on the boundary surface $\partial P_{\mathbf{X}}$ of $P_{\mathbf{X}}$ will produce a molecular surface $M_r(\mathbf{X})$ [5], Figure 1. Let d_w be the diameter of a water molecule and denote the molecular surface $M_{\frac{d_w}{2}}(\mathbf{X})$ as $M_{\mathbf{X}}$.

In general, any closed surface $S \subset \mathsf{R}^3$ will divide R^3 into three parts, $\mathsf{R}^3 = \Omega_S \cup S \cup \Omega'_S$, where Ω_S is a bounded domain and Ω'_S an un-bounded domain. Ω_S and Ω'_S have a common boundary, $\partial\Omega_S = \partial\Omega'_S = S$. If $M_{\mathbf{X}}$ has multiple connected components S_i, $1 \le i \le m$, such that S_1 is the largest component, i.e., all other components of $M_{\mathbf{X}}$ are contained in Ω_{S_1} (this is the case that $P_{\mathbf{X}}$ has $m-1$ cavities Ω_{S_i}, $i = 2, \cdots, m$, each is large enough to hold a water molecule), then denote $\Omega_{\mathbf{X}} = \Omega_{S_1} \cap (\cap_{i\neq 1}\Omega'_{S_i})$ and $\Omega'_{\mathbf{X}} = \Omega'_{S_1} \cup (\cup_{i\neq 1}\Omega_{S_i})$. If $M_{\mathbf{X}}$ is connected, then it is a closed surface. Thus, we always have

$$\mathsf{R}^3 = \Omega_{\mathbf{X}} \cup M_{\mathbf{X}} \cup \Omega'_{\mathbf{X}}, \quad \partial\Omega_{\mathbf{X}} = \partial\Omega'_{\mathbf{X}} = M_{\mathbf{X}}, \tag{1}$$

and $P_{\mathbf{X}} \subset \overline{\Omega_{\mathbf{X}}} = \Omega_{\mathbf{X}} \cup M_{\mathbf{X}}$.

For any compact (closed and bounded) set $U \subset \mathsf{R}^3$, let $\mathrm{dist}(\mathbf{x}, U) = \min_{\mathbf{z}\in U} |\mathbf{x} - \mathbf{z}|$ be the distance between the point $\mathbf{x}$ and the subset U. Define

$$\mathsf{R}_{\mathbf{X}} = \{\mathbf{x} \in \mathsf{R}^3 : \mathrm{dist}(\mathbf{x}, M_{\mathbf{X}}) \le d_w\} \setminus \Omega_{\mathbf{X}}, \quad \mathsf{T}_{\mathbf{X}} = \Omega_{\mathbf{X}} \cup \mathsf{R}_{\mathbf{X}}.$$

The $R_{\mathbf{X}}$ is the first hydration shell surrounding $P_{\mathbf{X}}$, and $\mathsf{T}_{\mathbf{X}}$ is the thermodynamic system tailor made for the conformation $\mathbf{X}$. Define compact sets $P_{\mathbf{X}i} = \cup_{\mathbf{a}_j \in H_i} B(\mathbf{x}_j, r_j)$,

$$\mathsf{R}_{\mathbf{X}i} = \{\mathbf{x} \in \mathsf{R}_{\mathbf{X}} : \mathrm{dist}(\mathbf{x}, P_{\mathbf{X}_i}) \le \mathrm{dist}(\mathbf{x}, P_{\mathbf{X}} \setminus P_{\mathbf{X}i})\}, \mathrm{FIG2}$$

and $M_{\mathbf{X}i} = M_{\mathbf{X}} \cap \mathsf{R}_{\mathbf{X}i}$, $i = 1, \cdots, H$. We will allow water and electrons enter or leave $\mathsf{T}_{\mathbf{X}}$, so $\mathsf{T}_{\mathbf{X}}$ is an open system.

Interaction of a water molecule with an atom of H_i will gain a Gibbs energy μ_i, the chemical potential. Let ν_i be the average number of water molecules that can simultaneously touch $M_{\mathbf{X}i}$ in a unit area, then $\omega_i = \nu_i \mu_i$ is the

ISSN: 2347-3487

chemical potential per unit area of $M_{\mathbf{X}i}$. Moreover, since the curvature of $M_{\mathbf{X}}$ is uniformly bounded for all conformations of U, ω_i does not depend on $\mathbf{X}$. Similarly we define μ_e and ω_e to be the chemical potentials of per electron and per unit volume.

Theorem 1 *Let* U *be a monomeric globular protein with* M *atoms* $(\mathbf{a}_1,\cdots,\mathbf{a}_M)$ *and* $\mathbf{X}=(\mathbf{x}_1,\cdots,\mathbf{x}_M)$ *be a conformation. Then in the physiological environment* $\mathrm{E}n_{\mathbf{N}}$

$$G(\mathbf{X};\mathsf{U},\mathrm{E}n_{\mathbf{N}})=\mu_e N_e(\mathbf{X})+\sum_{i=1}^{H}\mu_i N_i(\mathbf{X})+\sum_{1\le A<B\le N}\frac{q_A q_B}{4\pi\varepsilon_0\,|\,\mathbf{x}_A-\mathbf{x}_B\,|},\quad(2)$$

where q_A is the electronic charge in the nucleus of $\mathbf{a}_A$, $N_e(\mathbf{X})$ and $N_i(\mathbf{X})$ mean numbers of electrons in $\mathsf{T}_{\mathbf{X}}$ and water molecules in $\mathsf{R}_{\mathbf{X}i}$ respectively. The geometric version is

$$G(\mathbf{X};\mathsf{U},\mathrm{E}n_{\mathbf{N}})=\omega_e V(\Omega_{\mathbf{X}})+d_w\omega_e A(M_{\mathbf{X}})+\sum_{i=1}^{H}\omega_i A(M_{\mathbf{X}i})+\sum_{1\le A<B\le M}\frac{q_A q_B}{4\pi\varepsilon_0\,|\,\mathbf{x}_A-\mathbf{x}_B\,|}.\quad(3)$$

Proof of Theorem 1: Since water molecules are very small comparing to U, we can apply the Born-Oppenheimer approximation, fix $P_{\mathbf{X}}$ and let all water molecules and electrons in $\mathsf{T}_{\mathbf{X}}$ move. Then we will apply the grand canonic ensemble of statistical mechanics to the open system $\mathsf{T}_{\mathbf{X}}$.

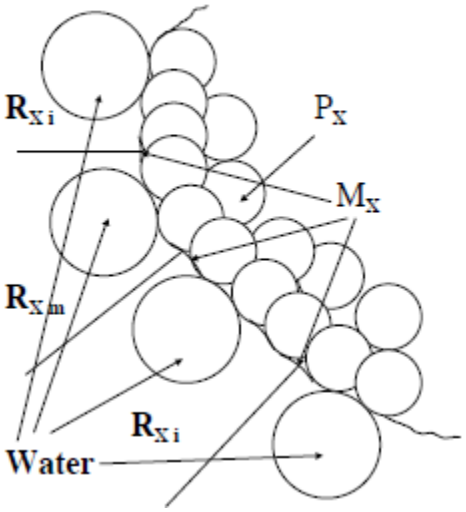

Figure 2: $\mathsf{R}_{\mathbf{X}i}$ **may not be connected.**

A water molecule is treated as a single particle centered at the oxygen nuclear position $\mathbf{w}\in\mathsf{R}^3$, and the covalent bonds in it are fixed. In the Born-Oppenheimer approximation, only the conformation $\mathbf{X}$ is fixed, all particles, water molecules and electrons in $\mathsf{T}_{\mathbf{X}}$, are moving.

Let $\mathbf{W}=(\mathbf{w}_1,\cdots,\mathbf{w}_i,\cdots,\mathbf{w}_N)\in\mathsf{R}^{3N}$ be the nuclear centers of water molecules in $\mathsf{R}_{\mathbf{X}}$ and $\mathbf{E}=(\mathbf{e}_1,\cdots,\mathbf{e}_i,\cdots,\mathbf{e}_L)\in\mathsf{R}^{3L}$ be electronic positions of all electrons in $\mathsf{T}_{\mathbf{X}}$. Then the Hamiltonian for the system $\mathsf{T}_{\mathbf{X}}$ is

$$\hat{H} = \hat{T} + \hat{V} = -\sum_{i=1}^{M} \frac{\hbar^2}{2m_i} \nabla_i^2 - \frac{\hbar^2}{2m_w} \sum_{i=1}^{N} \nabla_i^2 - \frac{\hbar^2}{2m_e} \sum_{i=1}^{L} \nabla_i^2 + \hat{V}(\mathbf{X}, \mathbf{W}, \mathbf{E}),$$

where m_i is the nuclear mass of atom $\mathbf{a}_i$, m_w and m_e are the masses of a water molecule and an electron; ∇_i^2 is Laplacian in corresponding R^3; and $V(\mathbf{X}, \mathbf{W}, \mathbf{E})$ is the potential.

We assume that a water molecule cannot occupy spaces in $P_{\mathbf{X}}$, thus by the design of $\mathsf{T}_{\mathbf{X}}$, any $\mathbf{w}_i$ in $\mathbf{W}$ belongs to $\mathsf{R}_{\mathbf{X}}$. Consider all possible numbers N_i of water molecules contained in $\mathsf{R}_{\mathbf{X}i}$, $0 \le \sum_{i=1}^{H} N_i = N < \infty$. Let $M_0 = 0$ and $M_i = \sum_{j \le i} N_j$ and $\mathbf{W}_i = (\mathbf{w}_{M_{i-1}+1}, \cdots, \mathbf{w}_{M_{i-1}+j}, \cdots, \mathbf{w}_{M_i}) \in \mathsf{R}_{\mathbf{X}i}^{N_i}$, $1 \le i \le H$, and $\mathbf{W} = (\mathbf{W}_1, \mathbf{W}_2, \cdots, \mathbf{W}_N) \in \prod_{i=1}^{H} \mathsf{R}_{\mathbf{X}i}^{N_i}$ denote the nuclear positions of water molecules in $\mathsf{R}_{\mathbf{X}}$. Similarly, consider all possible numbers $0 \le N_e < \infty$ of electrons in $\mathsf{T}_{\mathbf{X}}$. Let $\mathbf{E} = (\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_{N_e}) \in \mathsf{T}_{\mathbf{X}}^{3N_e}$ denote their nuclear positions. Denote $\mathbf{N} = (N_1, \cdots, N_H, N_e)$.

The potential $V(\mathbf{X}, \mathbf{W}, \mathbf{E})$ can be decomposed as:

$$V(\mathbf{X}, \mathbf{W}, \mathbf{E}) = V(\mathbf{X}) + V_{\mathbf{X}}(\mathbf{W}, \mathbf{E}), \tag{4}$$

where $V(\mathbf{X})$ is independent of $\mathbf{W}$ and $\mathbf{E}$. The Born-Oppenheimer approximation has the Hamiltonian (here we use the definition in [6]):

$$\hat{H}_X = -\frac{\hbar^2}{2} \left\{ \frac{1}{m_w} \sum_{j=1}^{M_H} \nabla_j^2 + \frac{1}{m_e} \sum_{\nu=1}^{N_e} \nabla_\nu^2 \right\} + \hat{V}_{\mathbf{X}}(\mathbf{W}, \mathbf{E}). \tag{5}$$

The eigenfunctions $\psi_i^{\mathbf{X},\mathbf{N}}(\mathbf{W}, \mathbf{E}) \in L_0^2(\prod_{i=1}^{H} \mathsf{R}_{\mathbf{X}i}^{N_i} \times \mathsf{T}_{\mathbf{X}}^{N_e}) = \mathsf{H}_{\mathbf{X},\mathbf{N}}$, comprise an orthonormal basis of $\mathsf{H}_{\mathbf{X},\mathbf{N}}$. Denote their eigenvalues as $E_{\mathbf{X},\mathbf{N}}^i$, then $\hat{H}_{\mathbf{X}} \psi_i^{\mathbf{X},\mathbf{N}} = E_{\mathbf{X},\mathbf{N}}^i \psi_i^{\mathbf{X},\mathbf{N}}$.

Since $\mathbf{N}$ varies, we can adopt the grand canonic ensemble. The grand canonic density operator is given as ([7] and [8])

$$\hat{\rho}_{\mathbf{X}} = \exp\left\{ -\beta \left[\hat{H}_{\mathbf{X}} - \sum_{i=1}^{H} \mu_i \hat{N}_i - \mu_e \hat{N}_e - \Phi(\mathbf{X}) \right] \right\}.$$

where $\beta = 1/(kT)$. The grand partition function is

$$\exp[-\beta \Phi(\mathbf{X})] = \sum_{i,\mathbf{N}} e^{-\beta \left[E_{\mathbf{X},\mathbf{N}}^i - \sum_{j=1}^{H} \mu_j N_j - \mu_e N_e \right]}.$$

According to [7], the entropy $S(\mathbf{X}) = S(\mathsf{T}_{\mathbf{X}})$ is

$$S(\mathbf{X}) = -k \mathrm{Trace}(\hat{\rho}_{\mathbf{X}} \ln \hat{\rho}_{\mathbf{X}}) = -k \langle \ln \hat{\rho}_{\mathbf{X}} \rangle = \frac{1}{T} \left[\langle \hat{H}_{\mathbf{X}} \rangle - \Phi(\mathbf{X}) - \sum_{i=1}^{H} \mu_i \langle \hat{N}_i \rangle - \mu_e \langle \hat{N}_e \rangle \right]. \tag{6}$$

The term $\Phi(\mathbf{X})$ is the grand canonic potential ϕ in [8] and the macroscopic potential in [7], it is a state function with variables $T, V, \mu_1, \cdots, \mu_H$, and μ_e. By the general thermodynamic equations [8]:

$$\mathrm{d}\Phi(\mathbf{X}) = -S\mathrm{d}T - P\mathrm{d}V - \sum_{i=1}^{H} N_i \mathrm{d}\mu_i - N_e \mathrm{d}\mu_e, \text{and}$$

ISSN: 2347-3487

$$\lambda\Phi(\mathbf{X}) = \Phi(\mathbf{X})(T, \lambda V, \mu_1, \cdots, \mu_H, \mu_e), \text{we see that}$$

$$\Phi(\mathbf{X})(T, V, \mu_1, \cdots, \mu_H, \mu_e) = -PV(\mathbf{X}), \tag{7}$$

where $V(\mathbf{X}) = V(\mathsf{T}_\mathbf{X})$ is the volume of $\mathsf{T}_\mathbf{X}$.

We denote $\langle \hat{N}_i \rangle = N_i(\mathbf{X})$ the mean numbers of water molecules in $\mathsf{R}_{\mathbf{X}i}$, $1 \le i \le H$, and $\langle \hat{N}_e \rangle = N_e(\mathbf{X})$ the mean number of electrons in $\mathsf{T}_\mathbf{X}$. By (6) and (7), we have

$$\langle \hat{H}_\mathbf{X} \rangle + PV(\mathbf{X}) - TS(\mathbf{X}) = \sum_{i=1}^{H} \mu_i N_i(\mathbf{X}) + \mu_e N_e(\mathbf{X}). \tag{8}$$

The mean $\langle \hat{H}_\mathbf{X} \rangle$ contains all energies of $\mathsf{T}_\mathbf{X}$ except the Coulomb potential $V(\mathbf{X})$. Thus the internal energy $U(\mathbf{X})$ is

$$U(\mathbf{X}) = U(\mathsf{T}_\mathbf{X}) = \langle \hat{H}_\mathbf{X} \rangle + V(\mathbf{X}). \tag{9}$$

In general, $G = U + PV - TS$. By (8), (9) and

$$V(\mathbf{X}) = \sum_{1 \le A < B \le N} \frac{q_A q_B}{4\pi\varepsilon_0 |\mathbf{x}_A - \mathbf{x}_B|},$$

$$G(\mathbf{X}; \mathsf{U}, \mathrm{En}_\mathbf{N}) = \sum_{i=1}^{H} \mu_i N_i(\mathbf{X}) + \mu_e N_e(\mathbf{X}) + V(\mathbf{X})$$

$$= \mu_e N_e(\mathbf{X}) + \sum_{i=1}^{H} \mu_i N_i(\mathbf{X}) + \sum_{1 \le A < B \le N} \frac{q_A q_B}{4\pi\varepsilon_0 |\mathbf{x}_A - \mathbf{x}_B|},$$

that is exactly (2).

Since every water molecule in $\mathsf{R}_{\mathbf{X}i}$ has contact with the surface $M_{\mathbf{X}i}$, $N_i(\mathbf{X})$ is proportional to the area $A(M_{\mathbf{X}i})$. Therefore, there are $\nu_i > 0$, such that

$$\nu_i A(M_{\mathbf{X}i}) = N_i(\mathbf{X}), \quad 1 \le i \le H. \tag{10}$$

Let $p_{\mathbf{X},\mathbf{N}}(\mathbf{x})$ be the electronic density distribution for $\mathbf{N} = (N_e, N_1, \cdots, N_H)$. By [2], $N_e = \int_{\mathsf{T}_\mathbf{X}} p_{\mathbf{X},\mathbf{N}}(\mathbf{x})\mathrm{d}\mathbf{x}$. There is a $\nu_{\mathbf{X},\mathbf{N}} > 0$ such that $N_e = \int_{\mathsf{T}_\mathbf{X}} p_{\mathbf{X},\mathbf{N}}(\mathbf{x})\mathrm{d}\mathbf{x} = \nu_{\mathbf{X},\mathbf{N}} V(\mathsf{T}_\mathbf{X})$. By the definition of $\mathsf{T}_\mathbf{X}$ and $\Omega_\mathbf{X}$, we have roughly $V(\mathsf{T}_\mathbf{X} \backslash \Omega_\mathbf{X}) = d_w A(M_\mathbf{X})$. Then taking the mean we have

$$N_e(\mathbf{X}) = \langle \hat{N}_e \rangle = \langle \hat{\nu}_{\mathbf{X},\mathbf{N}} V(\mathsf{T}_\mathbf{X}) \rangle = \langle \hat{\nu}_{\mathbf{X},\mathbf{N}} \rangle V(\mathsf{T}_\mathbf{X})$$

$$= \nu_e V(\mathsf{T}_\mathbf{X}) = \nu_e [V(\Omega_\mathbf{X}) + V(\mathsf{T}_\mathbf{X} \backslash \Omega_\mathbf{X})] = \nu_e V(\Omega_\mathbf{X}) + \nu_e d_w A(M_\mathbf{X}). \tag{11}$$

Substitute (10) and (11) into (2), letting $\omega_e = \nu_e \mu_e$, $\omega_i = \nu_i \mu_i$, we derive (3). Theorem 1 is proved.

$G(\mathbf{X};\mathsf{U},\mathrm{En}_{\mathbf{N}})$ were derived in [9, 10], only without the Coulomb potential $V(\mathbf{X})$. That is because the original definition of Born-Oppenhemer [11] was applied instead of the one in [6], though mathematically they are equivalent, physical explanations of internal energy are different.

To derive $G(\mathbf{X};\mathsf{U},\mathrm{En}_{\mathbf{N}})$, we have to take a single conformation $\mathbf{X}$ and its immediate environment to form a thermodynamic system $\mathsf{T}_{\mathbf{X}}$ tailor made for $\mathbf{X}$, against the intuition that statistics only deals with ensembles of conformations. Indeed we also used ensemble, but ours is ensemble of water molecules and electrons, not conformations. The emphasizing is on the interactions between a single conformation $\mathbf{X}$ and its immediate environment. The interaction is expressed via the space distribution of hydrophobicity classes H_1, $\cdots$, H_H on $P_{\mathbf{X}}$. The space distribution is a discretizing of the observable physical quantity $p_{\mathbf{X}}$, the electronic distribution of electrons. Those hydrophobicity classes H_i come from the general knowledge of amino acids. This single molecule treatment also emphases the 3-dimensional geometric shape $P_{\mathbf{X}}$ of $\mathbf{X}$, such that $\mathbf{X}$ is no longer a structureless point of a phase space of tremendous huge dimensions.

After over 20 years of single molecule experiments, see for example, [12, 13], the emerging of a single molecule theory of protein folding should be anticipated. In fact, theoretically directly applying fundamental physical laws to study protein folding is overdue for too long. And, against claims that protein folding is a practical field that does not need theory, it is much in need and need urgently. For example, because of $G(\mathbf{X};\mathsf{U},\mathrm{En}_{\mathbf{N}})$ is not known, *ab initio* structure prediction is lagged behind homologous prediction. With the analytic $G(\mathbf{X};\mathsf{U},\mathrm{En}_{\mathbf{N}})$, *ab initio* structure prediction becomes a pure mathematical problem of finding minimizers $\mathbf{X}_{\mathbf{N}} \in \mathsf{R}^{3M}$ of an analytic function $G(\cdot;\mathsf{U},\mathrm{En}_{\mathbf{N}}):\mathsf{R}^{3M} \to \mathsf{R}$

$$G(\mathbf{X}_{\mathbf{N}};\mathsf{U},\mathrm{En}_{\mathbf{N}}) = \min_{\mathbf{X}\in\mathsf{R}^{3M}} G(\mathbf{X};\mathsf{U},\mathrm{En}_{\mathbf{N}}), \tag{12}$$

if $\mathbf{X}_{\mathbf{N}}$ is a global minimizer. If $\mathbf{X}_{\mathbf{N}}$ is only a local minimizer, then it is still stable and satisfies

$$\nabla G(\mathbf{X}_{\mathbf{N}};\mathsf{U},\mathrm{En}_{\mathbf{N}}) = . \tag{13}$$

Moreover, in solving (12) we can use the rotatable dihedral angles $\Phi_{\mathbf{X}} = (\phi_1,\cdots,\phi_D)$ as variables, $G(\mathbf{X};\mathsf{U},\mathrm{En}_{\mathbf{N}}) = G(\Phi_{\mathbf{X}};\mathsf{U},\mathrm{En}_{\mathbf{N}})$. The advantages are that all bond lengths and angles are kept invariant, see [9]. Formulas of the gradient $\nabla G(\Phi_{\mathbf{X}};\mathsf{U},\mathrm{En}_{\mathbf{N}})$ are integrals on the molecular surface $M_{\mathbf{X}}$, see [9]. They are integrable and with packages of molecular surface, for example, [14], can be calculated precisely.

Lacking of theoretical guiding caused decades of misconceptions of protein folding problem. For example, the protein folding problem is artificially split into three parallel problems: 1. Folding code; 2. Structure prediction; 3. Kinetic process [15, 16]. They are treated separately as if no intrinsic relations could unify them. Absent of $G(\mathbf{X};\mathsf{U},\mathrm{En}_{\mathbf{N}})$ is the reason. In fact, code of protein folding in the meaning of the universal genetic code, does not exist [16]. Instead of code, the Gibbs free energy formula that governs the protein folding according the second law of thermodynamics, should be pursued. It not only provides physical basis and mathematical tools for *ab initio* structure prediction, but its gradient, the force $-\nabla G(\mathbf{X};\mathsf{U},\mathrm{En}_{\mathbf{N}})$ also gives a way of applying fundamental physical law to the kinetics of protein folding. With $G(\mathbf{X};\mathsf{U},\mathrm{En}_{\mathbf{N}})$ and $-\nabla G(\mathbf{X};\mathsf{U},\mathrm{En}_{\mathbf{N}})$, the three parts of protein folding problem will be treated uniformly.

References

[1] C. B. Anfinsen, Principles that govern the folding of protein chains. *Science*, **181**, 223 (1973).

[2] R. F. W. Bader, *Atoms in Molecules: A Quantum Theory* (Oxford: Clarendon Press, 1990).

[3] C. Tanford and J. Reynolds, *Nature's Robots: A History of Proteins* (Oxford University Press, 2001).

[4] E. Eisenberg and A. D. McLachlan, Solvation energy in protein folding and binding. *Nature*, **319**, 199 (1986).

[5] F. M. Richards, Areas, volumes, packing, and protein structure. *Ann. Rev. Biophys. Bioeng.*, **6**, 151 (1977).

[6] P. D. Haynes, Linear-scaling methods in *ab initio* quantum-mechanical calculations. A dissertation submitted for the degree of Doctor of Philosophy at the University of Cambridge. (1998).

[7] W. G. Greiner, N. S. Neise, and H. Stöker, *Thermodynamics and Statistical Mechanics* (New York, Berlin, ... Spriger-Verlag, 1994).

[8] X. Dai, *Advanced Statistical Physics* (Shanghai: Fudan University Press, 2007).

[9] Y. Fang, Gibbs free energy formula for protein folding. In: *Thermodynamics - Fundamentals and Its Application in Science*, edited by Morales-Rodriguez, R. Chapter 3. http://www.intechopen.com/books/thermodynamics-fundamentals-and-its-application-in-science

[10] Y. Fang, Ben-Naim's "pitfall": Don Quixote's windmill. *Open Journal of Biophysics*, **3**, 13 (2013). http://www.scirp.org/journal/ojbiphy)

[11] M. Born and J. R. Oppenheimer, On the quantum theory of molecules. *Annalen der Physik*, (in German) **389**(80), 457 (1927).

[12] D. A. Schafer, J. Gelles, M. P. Sheetz, and R. Landick, Trancription by single molecules of RNA polymerase observed by light microscopy. *Nature*, **352**, 444 (1991).

[13] A. Borgia, P. M. Williams, and J. Clarke, Single-molecule studies of protein folding. *Annu. Rev. Biochem*, **77**, 101 (2008).

[14] M. L. Connolly, Molecular Surfaces. http://www.biohedron.com/

[15] K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl, The Protein Folding Problem. *Annu. Rev. Biophys.* **37**, 289 (2008).

[16] A. Ben-Naim, *The Protein Folding Problem and Its Solutions* (World Scientific, New Jersey, London, Singapore, Beijing, Shanghai, Hong Gang, Taipei, Chennai, 2013).