Scientific Research Publishing

Journal of Biosciences and Medicines, 2024, 12, 311-335
https://www.scirp.org/journal/jbm
ISSN Online: 2327-509X
ISSN Print: 2327-5081

Applications of a Single Molecule Theory of Protein Dynamics

Yi Fang

School of Mathematics, Jilin University, Changchun, China
Email: yi.fang3@gmail.com

How to cite this paper: Fang, Y. (2024) Applications of a Single Molecule Theory of Protein Dynamics. *Journal of Biosciences and Medicines*, **12**, 311-335.
https://doi.org/10.4236/jbm.2024.126025

Received: May 11, 2024
Accepted: June 25, 2024
Published: June 28, 2024

Abstract

A single molecule theory for protein dynamics has been developed since 2012. It consists of the concepts of conformational Gibbs free energy function (CGF) and single molecule thermodynamic hypothesis (STH) that claims that all stable conformations are (local or global) minimizers of CGF. These are enough to give a unified explanations and mechanisms to many aspects of protein dynamics such as protein folding; allostery; denaturation; and intrinsically disordered proteins. Formulas of CGF in water environment had been derived via quantum statistics. Applications of them to soluble proteins are: docking Gibbs free energy difference formula and a practical way to search better docking site; single molecule binding affinity; predicting and explaining why structures of a monomeric globular protein looks like a globule and is tightly packed with a hydrophobic core; a representation of the hydrophobic effect; and a wholistic view to structures of water soluble proteins.

Keywords

Folding, Denaturation, Binding, Post-Binding Deformation, Allostery, Dynamic Second Law of Thermodynamics

1. Introduction

Up to the late 1970s, biochemistry had only dealt with molecular ensembles for which the laws of thermodynamics are readily applicable.

Ensemble methods have played and are still playing central roles in biological research, and have produced wonderful results such as the geometric principle of protein functioning and proteins having native structures, [1] and [2]. Anfinsen's Thermodynamic Hypothesis was inferred from experiments and observations on molecular ensembles, which claims that "This hypothesis states that the three-dimensional structure of a native protein in its normal physiological mi-

lieu (solvent, pH, ionic strength, presence of other components such as metal ions or prosthetic groups, temperature, and other) is the one in which the Gibbs free energy of the whole system is lowest; that is, that the native conformation is determined by the totality of interatomic interactions and hence by the amino acid sequence, in a given environment." [3].

In ensemble experiments, single molecule behaviour can only be inferred from the average of the ensemble. Averaging methods cannot really tell us the individual molecule's dynamics.

But the reality in biology is that in natural biological environment, only a few molecules of the same biological macromolecules are involved in any particular interaction. For example, according to [4], for some yeast proteins, there are only 60 molecules per cell. Thus to follow nature, experiments, measuring, and observing of biological macromolecules should take single molecule methods. Indeed, the introduction of [5] is entitled as "Molecular biophysics at the twenty-first century: from ensemble measurements to single-molecule detection".

Single molecule experiments are well developed now, single molecule theory lags behind. To solve the protein folding problem, the author has tried to develop a single molecule theory of protein dynamics based on two concepts, single molecule conformational Gibbs free energy function (CGF) and single molecule thermodynamic hypothesis (STH) that claims that all stable conformations are (local or global) minimizers of CGF, [6]-[13].

2. Method

Although a cell is crowded with various molecules, ions, etc., in protein folding process at any moment there are just a few molecules of the same protein, as stated in [14], "Folding generally involves only one molecule at a time, working, at least in most cases, without the aid of any other molecular actors except a suitable solvent. So no fancy biology needs to be invoked—chaperones, which after all consume valuable ATP, are actually used quite sparingly *in vivo*." Thus, in dealing with protein dynamics we may imaging that protein molecules fold in their physiological environment independently to one another.

We can make a mental experiment that a single protein molecule folds in certain environment, its conformation changes from the initial conformation to the final, native structure, leaving a folding path consisting of a series conformations. Each of the conformations in the folding path possesses a Gibbs free energy whose value also depends on the environment such that the final, native structure has the minimum Gibbs free energy among all conformations in the folding path. These are the ideas of conformational Gibbs free energy function and single molecule thermodynamic hypothesis. The formal description is as follows.

Suppose that $\mathfrak{U}$ is a molecule consisting of n atoms $(\mathbf{a}_1,\cdots,\mathbf{a}_n)$. A conformation of $\mathfrak{U}$ is a point in the 3n-dimensional Euclidean space $\mathbb{R}^{3n}$ and denoted as $\mathbf{R}=(\mathbf{r}_1,\cdots,\mathbf{r}_n)\in\mathbb{R}^{3n}$, where $\mathbf{r}_i=(x_i,y_i,z_i)\in\mathbb{R}^3$ is the nuclear centre of the atom $\mathbf{a}_i$. When talking conformations we assume all covalent bonds and bond angles are correctly formed. There are standard bond lengths and angles

for a molecule $\mathfrak{U}$ which should be respected in any conformation $\mathbf{R}$ of $\mathfrak{U}$. Thus, we require that a conformation $\mathbf{R}=(\mathbf{r}_1,\cdots,\mathbf{r}_n)\in\mathbb{R}^{3n}$ of $\mathfrak{U}$ satisfying the steric conditions below.

$$\begin{cases} |b_{ij}-r_{ij}|\le\delta_{ij} & \text{if } \mathbf{a}_i \text{ and } \mathbf{a}_j \text{ are bonded,} \\ r_{ij}>r_i+r_j & \text{if } \mathbf{a}_i \text{ and } \mathbf{a}_j \text{ are not bonded,} \\ |\gamma_{ij,ik}-\alpha_{ij,ik}|\le\beta_{ij,ik} & \text{if } \mathbf{a}_i \text{ bonds with both } \mathbf{a}_j \text{ and } \mathbf{a}_k. \end{cases} \quad (1)$$

where $r_{ij}=|\mathbf{r}_i-\mathbf{r}_j|=\left[(x_i-x_j)^2+(y_i-y_j)^2+(z_i-z_j)^2\right]^{\frac{1}{2}}$ is the distance between $\mathbf{r}_i$ and $\mathbf{r}_j$, r_i and r_j are the van der Waals radii of $\mathbf{a}_i$ and $\mathbf{a}_j$, δ_{ij} and $\beta_{ij,ik}$ are small positive constants, $b_{ij}<r_i+r_j$ are the standard bond lengths between $\mathbf{a}_i$ and $\mathbf{a}_j$, $\alpha_{ij,ik}$ is the standard bond angle if there are covalent bonds between $\mathbf{a}_i$ and $\mathbf{a}_j$ and also between $\mathbf{a}_i$ and $\mathbf{a}_k$, $\gamma_{ij,ik}$ are bond angles measured in $\mathbf{R}$. The set of all conformations of $\mathfrak{U}$ is denoted as $\mathfrak{X}_{\mathfrak{U}}\subset\mathbb{R}^{3n}$.

Let a molecule $\mathfrak{U}$ be situated in an environment $\mathfrak{E}$ (solvent, pH, ionic strength, temperature, pressure, etc.). Each conformation $\mathbf{R}\in\mathfrak{X}_{\mathfrak{U}}$ occupies a space $V_{\mathbf{R}}$ in $\mathbb{R}^3$. Adding one layer of environment particles surrounding $V_{\mathbf{R}}$, we get a tiny open thermodynamic system $\mathfrak{S}_{\mathbf{R}}\subset\mathbb{R}^3$ tailor made for the conformation $\mathbf{R}$ ("open" means that particles of $\mathfrak{E}$ can enter and leave the system). The Gibbs free energy $G(\mathfrak{S}_{\mathbf{R}})$ of $\mathfrak{S}_{\mathbf{R}}$ can be denoted as the value of a function $G(\mathbf{R};\mathfrak{U},\mathfrak{E})$ at $\mathbf{R}$. $G(\mathbf{R};\mathfrak{U},\mathfrak{E})$ is a single molecule conformational Gibbs free energy function (CGF), a function whose variables are the conformations **R**s. $G(\mathbf{R};\mathfrak{U},\mathfrak{E})$ also has two groups of parameters $\mathfrak{U}$ and $\mathfrak{E}$, the molecule and the environment.

Note that for different environments $\mathfrak{E}_1$ and $\mathfrak{E}_2$, $G(\mathbf{R},\mathfrak{U},\mathfrak{E}_1)$ and $G(\mathbf{R};\mathfrak{U},\mathfrak{E}_2)$ are different functions. For example, if $\mathbf{R}_1$ and $\mathbf{R}_2$ are two different conformations of $\mathfrak{U}$, comparing $G(\mathbf{R}_1;\mathfrak{U},\mathfrak{E}_1)$ and $G(\mathbf{R}_2,\mathfrak{U},\mathfrak{E}_2)$ does not make sense. But comparing $G(\mathbf{R};\mathfrak{U},\mathfrak{E}_1)$ and $G(\mathbf{R};\mathfrak{U},\mathfrak{E}_2)$ does make sense, it shows the environment influence to the same conformation $\mathbf{R}$. Similarly, let $\mathfrak{U}_1$ and $\mathfrak{U}_2$ be two molecules, in general we should not compare $G(\mathbf{R};\mathfrak{U}_1,\mathfrak{E})$ and $G(\mathbf{R};\mathfrak{U}_2,\mathfrak{E})$, since to begin with, $\mathfrak{X}_{\mathfrak{U}_1}$ and $\mathfrak{X}_{\mathfrak{U}_2}$ are different.

The single molecule thermodynamic hypothesis (STH), is a single molecule version of Anfinsen's Thermodynamic Hypothesis (ATH). It claims that in the environment $\mathfrak{E}$, all stable conformations of $\mathfrak{U}$ must be (local or global) minimizers of the CGF $G(\mathbf{R};\mathfrak{U},\mathfrak{E})$.

For small monomeric globular proteins (100 to 400 residues) many believe that the native structure $\mathbf{R}_{\mathfrak{U}}$ of $\mathfrak{U}$ is the global minimizer of $G(\mathbf{R};\mathfrak{U},\mathfrak{W})$ where $\mathfrak{W}$ is the aqueous solvent (water) environment, *i.e.*,

$$G(\mathbf{R}_{\mathfrak{U}};\mathfrak{U},\mathfrak{W})\le G(\mathbf{R};\mathfrak{U},\mathfrak{W}), \text{ for all } \mathbf{R}\in\mathfrak{X}_{\mathfrak{U}}. \quad (2)$$

In 2011, the concept of CGF was thought as absurd, see [15] in which it was argued that such a function cannot exist and is the biggest pitfall of ATH. In fact, formula of the CGF, $G(\mathbf{R};\mathfrak{U},\mathfrak{W})$, has been derived via quantum statistics,

$$G(\mathbf{R};\mathfrak{U},\mathfrak{W}) = U(\mathbf{R}) + \mu_e N_e(\mathbf{R}) + \sum_{i=1}^{L} \mu_i N_i(\mathbf{R}), \tag{3}$$

where $U(\mathbf{R})$ is the intra-conformational potential energy depending only on the conformation $\mathbf{R}$; $N_e(\mathbf{R})$ is the mean number of electrons in the tiny open thermodynamic system $\mathfrak{S}_{\mathbf{R}}$ and $\mu_e > 0$ is the chemical potential of an electron; $N_i(\mathbf{R})$'s are mean numbers of first layer water molecules surrounding $V_{\mathbf{R}}$ which are positioned nearby moieties of $\mathfrak{U}$ that having hydrophobicity level i, $1 \le i \le L$, such water molecules then have chemical potentials μ_i's. Thus, the chemical potential of a water molecule in $\mathfrak{S}_{\mathbf{R}}$ depends on the water molecule's position. All hydrophobic moieties have positive μ_i's, all hydrophilic moieties have negative μ_i's. $L \ge 1$ in general, if $\mathfrak{U}$ is a protein then $L > 1$ since proteins are amphiphiles having at least two hydrophobicity classes. Occasionally, depending on conformation $\mathbf{R}$, moieties may change from hydrophilic to hydrophobic. For example, when two polar or charged moieties formed a hydrogen bond or neutralized their charges due to their close up in $\mathbf{R}$, then the two moieties should be denoted as hydrophobic in $\mathbf{R}$. These intramolecular hydrogen bonds and charge neutralizations are important elements of the intra-conformational potential function $U(\mathbf{R})$.

A geometric approximation of (3) is by an interface $\Sigma_{\mathbf{R}} \subset \mathfrak{S}_{\mathbf{R}}$ between $V_{\mathbf{R}}$ and water molecules in $\mathfrak{S}_{\mathbf{R}}$. $\Sigma_{\mathbf{R}}$ is a closed surface (no boundary) of genus zero (homeomorphic to a sphere). $\Sigma_{\mathbf{R}}$ is the boundary of a bounded domain $\Omega_{\mathbf{R}} \subset \mathbb{R}^3$ such that $V_{\mathbf{R}} \subset \bar{\Omega}_{\mathbf{R}} = \Omega_{\mathbf{R}} \cup \partial\Omega_{\mathbf{R}}$, where the boundary $\partial\Omega_{\mathbf{R}}$ of $\Omega_{\mathbf{R}}$ is just $\Sigma_{\mathbf{R}}$. All water molecules in $\mathfrak{S}_{\mathbf{R}}$ are contained in $\mathfrak{S}_{\mathbf{R}} \setminus \Omega_{\mathbf{R}}$ except that $V_{\mathbf{R}}$ has interior holes (contained in $\Omega_{\mathbf{R}}$) big enough to contain some water molecules, in this case, these interior water molecules are not counted in the one layer water molecules.

The geometric approximation of Formula (3) is as follows:

$$G(\mathbf{R},\Sigma_{\mathbf{R}};\mathfrak{U},\mathfrak{W}) = U(\mathbf{R}) + \omega_e V(\Omega_{\mathbf{R}}) + d_w \omega_e A(\Sigma_{\mathbf{R}}) + \sum_{i=1}^{L} \omega_i A(\Sigma_{\mathbf{R},i}), \tag{4}$$

$V(\Omega_{\mathbf{R}})$ and $A(\Sigma_{\mathbf{R}})$ are volume and area, $\omega_e > 0$ the electron chemical potential per volume; d_w the diameter of a water molecule; $A(\Sigma_{\mathbf{R},i})$ the area of the subsurface $\Sigma_{\mathbf{R},i} \subset \Sigma_{\mathbf{R}}$ covering all moieties of hydrophobicity level i that are exposed to water in $\mathbf{R}$, $\sum_{i=1}^{L} A(\Sigma_{\mathbf{R};i}) = A(\Sigma_{\mathbf{R}})$, ω_i's the chemical potentials per area.

The advantage of $G(\mathbf{R},\Sigma_{\mathbf{R}};\mathfrak{U},\mathfrak{W})$ in (4) is that it is differentiable, thus we can use its gradient $\nabla G(\mathbf{R},\Sigma_{\mathbf{R}};\mathfrak{U},\mathfrak{W})$ to discuss such as stable conformations and folding dynamics. For example, the negative gradient $-\nabla_{\mathbf{R}} G(\mathbf{R},\Sigma_{\mathbf{R}};\mathfrak{U},\mathfrak{W})$ will be the main folding force (Anfinsen stated in [3] that "This process (protein folding) is driven entirely by the free energy of conformation that is gained in going to stable, native structure."). The advantage of Formula (3) is that it is easy to grasp and apply to formulations, such as the docking Gibbs free energy difference in §3.4.

Formulas (3) and (4) are derived via quantum statistics in [6]-[9] and [11].

Note that it is always deriving Formula (3) first via quantum statistics, then approaching it by Formula (4). To fill in the argument in limited pages, preparation materials are just treated as well known. Some readers may feel difficult to understand the derivation. Detailed derivation, including all necessary preparations in thermodynamics, statistic mechanics, and quantum chemistry, etc., will appear in a coming monograph, the preparation and derivation will occupy over 70 pages.

In previous derivations, the intra-conformational potential $U(\mathbf{R})$ was presented only as the Coulomb's potential of nuclei $\sum_{i\neq j}\frac{q_i q_j e^2}{\varepsilon r_{ij}}$. This should be modified to also consider potentials involving electron cloud.

The term $\mu_e N_e(\mathbf{R})$ in (3) comes purely from quantum mechanics. Applying classical statistics will result in the same formula as (3) but without this electron quantum term. Term $\omega_e V(\Omega_{\mathbf{R}}) + d_w \omega_e A(\Sigma_{\mathbf{R}})$ in (4) comes from translating $\mu_e N_e(\mathbf{R})$ to $\omega_e V(\mathfrak{S}_{\mathbf{R}})$ via a quantum mechanics argument, then approximate $V(\mathfrak{S}_{\mathbf{R}})$ by $V(\Omega_{\mathbf{R}}) + d_w A(\Sigma_{\mathbf{R}})$.

The term $\sum_{i=1}^{L}\mu_i N_i(\mathbf{R})$ in (3) and its counterpart $\sum_{i=1}^{L}\omega_i A(\Sigma_{\mathbf{R},i})$ in (4) come from the surrounding water molecules, thus can be called the Gibbs free energy of aqueous solvent. It reduces the heavy calculations involving surrounding water molecules in molecular dynamic simulation to an easier to calculate boundary energy formula. It is called solvation free energy in literature. Molecular dynamics simulators pursued it but unsuccessful, only resulting in a no calculable integral formula that is not recognized as the solvation free energy but is called the potential of mean force or effective energy, see for example, [16] [17].

The occupied space $V_{\mathbf{R}}$ of a conformation $\mathbf{R}$ is determined by an electron wave function Φ in quantum mechanics. It is enveloped by the largest connected branch of the boundary $\partial V_{\mathbf{R}} = \{\mathbf{x}\in\mathbb{R}^3; |\Phi(\mathbf{x})|^2 = \varepsilon\}$, where $\varepsilon > 0$ is a small number.

To make the interface $\Sigma_{\mathbf{R}}$ easy to calculate, we replace $V_{\mathbf{R}}$ by a very good approximation, a bunch of overlapping balls $P_{\mathbf{R}} = \bigcup_{i=1}^{n} B(\mathbf{r}_i, r_i) \subset \mathbb{R}^3$, where $B(\mathbf{r}_i, r_i) \subset \mathbb{R}^3$ is the round ball centred at $\mathbf{r}_i$ with the van der Waals radius r_i of the atom $\mathbf{a}_i$. Figures on page 4 of [18], where $\partial V_{\mathbf{R}} = \{\mathbf{x}\in\mathbb{R}^3; |\Phi(\mathbf{x})|^2 = \varepsilon\}$, $\varepsilon = 0.001$ au, 1 au = a_0 = 0.529188 Å, show that $V_{\mathbf{R}}$ is so like a bunch of overlapping balls.

With $P_{\mathbf{R}}$ replacing $V_{\mathbf{R}}$, there are many choices of the interface $\Sigma_{\mathbf{R}}$, such as the van der Waals surface (it is just the outmost component of $\partial P_{\mathbf{R}}$); the solvent accessible surface defined by Lee and Richards in [19]; and the molecular surface defined by Richards in [20]. The first two are both bunches of overlapping spheres. The latter two surfaces are both generated by rolling a sphere of diameter d_w (the diameter of a water molecule) over $P_{\mathbf{R}}$ (or $V_{\mathbf{R}}$, only that we do not know exactly what is $V_{\mathbf{R}}$), see [21] and [22], if in case they are not connected, we take the outmost component of them as $\Sigma_{\mathbf{R}}$.

A question was often asked, what are the use of these concepts and formulations? This article discusses the applications of this single molecule theory, predictions and explanations to various aspects of protein dynamics and structures. Mechanisms of dynamic protein structure, in particular protein folding, binding, allostery, denaturation, all are predicted and explained by CGF and STH.

3. Discussions

A theory should be able to make verifiable predictions and explain various known phenomena, the more the better.

With the shift of point of view from ensemble to single molecule, the concept of CGF together with STH can explain phenomena such as allostery, protein denaturation, and intrinsically disordered proteins.

In §3.1, STH predicts that after binding there will be conformational change, thus allostery is result of post-binding deformation, i.e., after binding conformational change, either they have allosteric effect or not. Thus the mechanism of the second secret of life-allostery, is nothing but the dynamic second law of thermodynamics. Life follows the dynamic second law of thermodynamics.

In §3.2, we argue that by CGF and STH, denaturation and folding are both the process of reducing conformational Gibbs free energy to achieve a stable conformation, the only difference is that they happen in different environments.

In §3.3 we discuss intrinsically disordered proteins, pointing out they are caused by the ensemble point of view. In fact, by CGF and STH, each individual molecule will find a stable conformation. If in an ensemble of molecules of the same protein there are multiple stable conformations, we get intrinsically disordered proteins.

Even not knowing the accurate values of the chemical potentials prevents us to do the most important verifiable prediction, *ab initio* prediction of proteins' native structures, there are still other applications of the Formulas (3) and (4).

In §3.4, we derive a docking free energy difference formula ΔG_{ST} via (3). Unlike the quantitative task of structure prediction that needs accurate values of the chemical potentials, the docking free energy difference Formula (5) provides a qualitative but practical way to search would be docking sites that will be useful in finding new drugs. Finally we define the single molecule binding affinity such that the bigger the affinity the stronger the binding.

In §3.5, we apply (4) to predict the global geometric feathers of monomeric globular proteins, these predictions also explain the physics behind the well known global geometric features of globular proteins, they look like globules and are compactly packed, often, with hydrophobic cores in the structure interiors.

In §3.6 we argue that the dominate folding force is quantum force and hydrophobic force (aqueous force). We will show only a vanisingly tiny portion of polypeptide chains can be proteins. To be a monomeric globular protein, the entire polypeptide chain must be able to be arranged a delicate and balanced spacial arrangement to lower the intra-conformational potential $U(\mathbf{R})$. We also

argue that $U(\mathbf{R})$ cannot be dominate folding force.

In §3.7, we suggest a wholistic, top-down point of view of monomeric globular protein structures and list evidences supporting this view.

3.1. Demystify the Second Secret of Life: Mechanism of Allostery

Proteins work through binding to ligand.

3.1.1. Post-Binding Deformation

A dynamic version of STH, also a prediction, is: If a conformation $\mathbf{R}$ of $\mathfrak{U}$ is not stable in environment $\mathfrak{E}$, then it will spontaneous (or be forced by $-\nabla G(\mathbf{R};\mathfrak{U},\mathfrak{E})$) fold to a stable conformation $\mathbf{R}_1$ which must be a (local or global) minimizer of $G(\mathbf{R};\mathfrak{U},\mathfrak{E})$.

STH predicts that after binding there should be conformational change, or post-binding deformation, let $\mathbf{M}=(\mathbf{x}_1,\cdots,\mathbf{x}_n)\in\mathbb{R}^{3n}$ and $\mathbf{L}=(\mathbf{y}_1,\cdots,\mathbf{y}_m)\in\mathbb{R}^{3m}$ be stable conformations of $\mathfrak{M}$ and $\mathfrak{L}$ in environment $\mathfrak{E}$, it is very unlikely that $\mathbf{ML}=(\mathbf{x}_1,\cdots,\mathbf{x}_n,\mathbf{y}_1,\cdots,\mathbf{y}_m)\in\mathbb{R}^{3(n+m)}$ is also a stable conformation of the complex $\mathfrak{ML}$, therefore, STH predicts that it will fold until it achies a stable conformation $\mathbf{M}_f\mathbf{L}_f$ that is necessarily a (local or global) minimizer of $G(\mathbf{RQ};\mathfrak{ML},\mathfrak{E})$. In particular, $G(\mathbf{M}_f\mathbf{L}_f;\mathfrak{ML},\mathfrak{E})<G(\mathbf{ML},\mathfrak{ML},\mathfrak{E})$.

3.1.2. Allostery

Post-binding deformation was first noted in allostery, thus our prediction of it was already virified.

In ([23], p. 434) allostery is described as "In both enzymes and receptors, the inherent function of the protein (catalysis, binding), which occurs at a certain location (e.g., the active site in enzymes), can be modulated by binding of a ligand at a different location. This phenomenon is called 'allostery' and the site to which the modulating ligand binds is called allosteric site."

Allostery is the second secret of life, "Monod's recognition of this important biological role led to the historical description of allostery as 'the second secret of life', second only to the genetic code." [24].

3.1.3. Mechanism of Allostery

The mechanism of allostery is the mechanism of post-binding deformation. There are no allostery proteins, any protein after binding will have conformational change. For some proteins, post-binding deformations have the results of allostery. Post-binding deformations also plays important roles in proteins' functions, for example, catalysis.

3.1.4. Dynamic Second Law of Thermodynamics

The second law of thermodynamics says that in an open thermodynamic system $\mathfrak{S}$ surrounded by a heat bath of constant temperature and pressure, the Gibbs free energy will achieve the minimum at equilibrium. But our $\mathfrak{S}_\mathbf{R}$ is not fixed, changing conformation to reduce the Gibbs free energy will also change the system $\mathfrak{S}_\mathbf{R}$, *i.e.*, we are pursuing a system with minimum Gibbs free energy among

all systems $\mathfrak{S}_{\mathbf{R}}$, instead of pursuing minimum Gibbs free energy for a fixed system $\mathfrak{S}$. We will call this kind of minimization of Gibbs free energy the dynamic second law of thermodynamics. Protein folding and binding follow this dynamic second law of thermodynamics.

3.1.5. Life Follows Dynamic Second Law of Thermodynamics

Proteins' functionalities rely on specific binding, the specificity comes from proteins' native structures. Native structures and post-binding deformations come from spontaneous folding whose mechanism is the dynamic second law of thermodynamics. Since protein folding and allostery play central roles in protein functioning and regulations in almost all cellular processes [25], we may safely say that life follows the dynamic second law of thermodynamics.

3.2. Mechanism of Denaturation

All known denature phenomena are caused by changing an environment $\mathfrak{E}_N$ in which a protein $\mathfrak{U}$ has a native structure to a denaturation environment $\mathfrak{E}_D$. Many ways can change environment. "Acids and alkalies, salts of heavy metals, alcohol, ether and other organic solvents, concentrated urea and related compounds, heat, ultraviolet light, high pressure, shaking, supersonic, waves and even drying: all these can induce denaturation of the protein." [2].

Roughly speaking, before the change of environment, an ensemble of protein molecules was in a kind of equilibrium such that most of these molecules are in native structure in the environment $\mathfrak{E}_N$. When the environment finally become the denaturation environment $\mathfrak{E}_D$, according to the STH, all molecules in this ensemble will fold to stable conformations, these stable conformations are minimizers of $G(\mathbf{R};\mathfrak{U},\mathfrak{E}_D)$.

The difference of folding and denaturation is that after folding almost all molecules take the same stable conformation $\mathbf{R}_{\mathfrak{U}}$, a minimizer of $G(\mathbf{R};\mathfrak{U},\mathfrak{E}_N)$; but in denaturation molecules will take many different stable conformations, $\mathbf{R}_{\mathfrak{U}}$ may or may not be a minimizer of $G(\mathbf{R};\mathfrak{U},\mathfrak{E}_D)$.

For example, consider $\mathbf{R}_{\mathfrak{U}}$ is the native structure of a protein $\mathfrak{U}$ in its physiological environment $\mathfrak{E}_N=\mathfrak{W}$ that has constant temperature T. Now we change the environment by lifting T. In terms specific (per unit mass) internal energy e_i and entropy s_i, the per unit area chemical potentials ω_i in (4) can be expressed as $m(e_i-Ts_i)$, where m is the mass of a water molecule. Since it is always $s_i>0$, if we lift T sufficiently high, eventually all ω_i and ω_e will become negative. In particular, hydrophobic area $A(\Sigma_{\mathbf{R},h})=0$, hydrophilic area $A(\Sigma_{\mathbf{R},p})=A(\Sigma_{\mathbf{R}})$. Once all chemical potentials become negative, according (4), the larger the volume $V(\Omega_{\mathbf{R}})$ and area $A(\Sigma_{\mathbf{R}})$, the lower the Gibbs free energy $G(\mathbf{R},\Sigma_{\mathbf{R}};\mathfrak{U},\mathfrak{W})$, the conformation will become an extended random coil. Continue lifing T will break the molecule. Stop lifting T before breaking, many extended random coils will become local minimizers of $G(\mathbf{R},\Sigma_{\mathbf{R}};\mathfrak{U},\mathfrak{W})$.

This is actually the observed phenomenon as stated in ([26]: p. 13), "Indeed the biological macromolecules, proteins as well as the nucleic acids attain or-

dered structure that are important and perform their functions in the cells as structures that are stable under certain circumstances, in particular low temperatures but can be disintegrated, 'denaturised' to flexible, non-ordered randomised coils at higher temperatures."

Why in $\mathfrak{E}_D$, folding from a single initial conformation $\mathbf{R}_{\mathfrak{U}}$ can result in different stable conformations? Are stable conformations independent of initial conformations?

Actually, the sudden changes of the same natives structure to many different structures in the initial denaturation process of an ensemble of protein molecules is a phenomenon often happenes in nature, mathematical description of it is the **catastrophe theory** developed by Rene Thom [27], a definition of it is given in [28]: "Catastrophe theory is concerned with the mathematical modelling of sudden changes—so called 'catastrophes'—in the behaviour of natural systems, which can appear as a consequence of continuous changes of the system parameters." Indeed, from the physiological environment $\mathfrak{E}_N$ to the denaturation environment $\mathfrak{E}_D$, there must be a one family of environments $\mathfrak{E}_t$ connecting them. Although these environments are continues parameters in the CGFs $G(\mathbf{R};\mathfrak{U},\mathfrak{E}_t)$, catastrophe does happen so that various copies of the same native structure $\mathbf{R}_{\mathfrak{U}}$ suddenly changed to different conformations for different molecules and when the environment parameter reached the denaturation environment $\mathfrak{E}_D$, these changed conformations become different initial conformations.

3.3. Intrinsically Disordered Proteins

A large proportion of proteins are intrinsically disordered, "After half a century of structural studies on beautifully folded globular proteins, it is perhaps a shock to discover that up to some 40% of the proteins in the human proteome are estimated to be intrinsically disordered and become fully or partly structured on binding to binding partners in the cell.", [29]. A recent literature analysis showed that there are approximately 1150 non-redundant proteins in the list of validated intrinsically disordered proteins, [30].

By CGF and STH, in the level of single molecule, any molecule always has a stable conformation and it must be a (local or global) minimizer of the CGF. For proteins having native structures, these molecules' stable conformations are the same. Repeat testing in the same environment always get the same stable conformation. For intrinsically disordered proteins, there are multiple stable conformations, therefore, repeat testing in the same environment will get different stable conformations, or none at all. Thus, looking in an ensemble, they seem to have no structure at all in an averaging observation such as X-ray crystollgraphy.

The concepts of ordered and disordered proteins were introduced in the level of ensemble of molecules. It is because that all protein structure determination methods working on ensembles of proteins, except the cryo electron microscopy. For example, the X-ray crystallography experiments are performed to crystals of ensembles of protein molecules. In the crystal, protein molecules are in cells, the

whole structure is obtained by adding up informations of individual cells together to get an average conformation. If all cells have the same conformation, for example, their structures are parallel motions (congruence) to each other, then we obtain a native structure by adding up all cells' information. But if in different cells the molecules' structures are different (no longer congruent), then the adding up of these structures will be blurred, showing no structure at all. Therefore, the situation for disordered proteins is: looking at individual molecules in the ensemble, each has a stable conformation, collectively as an ensemble, no average structure at all.

3.4. Docking and Binding

Formulas in (3) and (4) open the door of investigating all protein dynamic phenomena by first principle. Although quantitative investigations have to wait the relative accuracy chemical potential values to be determined, we still can do qualitative work like the following.

3.4.1. Docking Free Energy Difference

Here we just consider binding in the aqueous solvent $\mathfrak{W}$. Binding starts with docking, a process of two molecules come to each other and remove water molecules between surfaces of molecules (note that conformations are invariant under orientational preserving congruence of $\mathbb{R}^3$, so we use the same notations $\mathbf{R}$ and $\mathbf{Q}$ to express the moving conformations). Let $S \subset \Sigma_{\mathbf{R}}$ and $T \subset \Sigma_{\mathbf{Q}}$ be the docking sets, *i.e.*, water molecules between them are removed such that $\mathbf{R}$ and $\mathbf{Q}$ (almost) touching each other along S and T to form a new combined conformation $\mathbf{RQ} = (\mathbf{x}_1, \cdots, \mathbf{x}_n, \mathbf{y}_1, \cdots, \mathbf{y}_m) \in \mathbb{R}^{3(n+m)}$ of the would be complex $\mathfrak{UW}$.

Let $\mathfrak{S}_{\mathbf{RQ}}$ be the open thermodynamic system tailor made for the conformation $\mathbf{RQ}$, *i.e.*, it is composed by $V_{\mathbf{RQ}} \subset \mathbb{R}^3$ plus one layer of water molecules. The interface $\Sigma_{\mathbf{RQ}}$ inside $\mathfrak{S}_{\mathbf{RQ}}$ is

$$\Sigma_{\mathbf{RQ}} = (\Sigma_{\mathbf{R}} \setminus S) \cup (\Sigma_{\mathbf{Q}} \setminus T).$$

Starting from $\mathbf{RQ}$ along *ST*, whether or not a complex $\mathfrak{UW}$ can be formed and be further folded to a stable conformation will depend on the docking Gibbs free energy difference

$$\Delta G_{ST} = G(\mathfrak{S}_{\mathbf{RQ}}) - G(\mathbf{R}; \mathfrak{U}, \mathfrak{W}) - G(\mathbf{Q}; \mathfrak{V}, \mathfrak{W}).$$

If $\Delta G_{S \cup T} < 0$, $\mathbf{RQ}$ has the chance to further fold to a stable conformation $\mathbf{R}_1\mathbf{Q}_1$ of the complex $\mathfrak{UW}$; if $\Delta G_{ST} > 0$, the conformation $\mathbf{RQ}$ will not stand, the two separate conformations $\mathbf{R}$ and $\mathbf{Q}$ will be more stable than it. In other words, if $\Delta G_{ST} > 0$, the complex $\mathfrak{UW}$ has no chance to stabilize along the site *ST*.

3.4.2. Docking Free Energy Difference Formula

Suppose that there are $N^{\mathbf{R}}$ water molecules in $\mathfrak{S}_{\mathbf{R}}$, $N^{\mathbf{Q}}$ molecules in $\mathfrak{S}_{\mathbf{Q}}$, etc. Docking along *ST* means that $N_r > 0$ water molecules covering *S* and *T* are

removed from the open thermodynamic system $\mathfrak{S}_{\mathbf{RQ}}$. Denote $G(\mathfrak{S}_{\mathbf{RQ}})$ as the Gibbs free energy of $\mathfrak{S}_{\mathbf{RQ}}$, remembering that if $\mathfrak{W}$ along ST is possible, it is just $G(\mathbf{RQ};\mathfrak{W},\mathfrak{W})$. By the formula in (3),

$$G(\mathfrak{S}_{\mathbf{RQ}}) = U(\mathbf{RQ}) + \mu_e N_e(\mathbf{RQ}) + \sum_{i=1}^{L} \mu_i N_i(\mathbf{RQ}),$$

$$G(\mathbf{R};\mathfrak{U},\mathfrak{W}) = U(\mathbf{R}) + \mu_e N_e(\mathbf{R}) + \sum_{i=1}^{L} \mu_i N_i(\mathbf{R}),$$

$$G(\mathbf{Q};\mathfrak{V},\mathfrak{W}) = U(\mathbf{Q}) + \mu_e N_e(\mathbf{Q}) + \sum_{i=1}^{L} \mu_i N_i(\mathbf{Q}).$$

Let the total number of water molecules in $\mathfrak{S}_{\mathbf{RQ}}$, $\mathfrak{S}_{\mathbf{R}}$, and $\mathfrak{S}_{\mathbf{Q}}$ be

$$N^{\mathbf{RQ}} = \sum_{i=1}^{L} N_i(\mathbf{RQ}), \quad N^{\mathbf{R}} = \sum_{i=1}^{L} N_i(\mathbf{R}), \quad N^{\mathbf{Q}} = \sum_{i=1}^{L} N_i(\mathbf{Q}),$$

then

$$N^{\mathbf{RQ}} = N^{\mathbf{R}} + N^{\mathbf{Q}} - N_r.$$

Since

$$N_e(\mathbf{RQ}) - N_e(\mathbf{R}) - N_e(\mathbf{Q}) = 10[N^{\mathbf{RQ}} - N^{\mathbf{R}} - N^{\mathbf{Q}}] = -10N_r,$$

we have

$$\begin{aligned} \Delta G_{ST} &= U(\mathbf{RQ}) - U(\mathbf{R}) - U(\mathbf{Q}) - 10\mu_e N_r \\ &+ \sum_{\mu_i>0} \mu_i [N_i(\mathbf{RQ}) - N_i(\mathbf{R}) - N_i(\mathbf{Q})] \\ &+ \sum_{\mu_i<0} \mu_i [N_i(\mathbf{RQ}) - N_i(\mathbf{R}) - N_i(\mathbf{Q})]. \end{aligned} \tag{5}$$

3.4.3. Searching for Potential Docking Sites

To make ΔG_{ST} negative, the best chance is that we make all terms in (5) negative or zero. Since $\mu_e > 0$, $-10\mu_e N_r < 0$. To make $\sum_{\mu_i>0} \mu_i [N_i(\mathbf{RQ}) - N_i(\mathbf{R}) - N_i(\mathbf{Q})]$ negative, there should be more $N_i(\mathbf{RQ}) - N_i(\mathbf{R}) - N_i(\mathbf{Q}) < 0$ than $N_i(\mathbf{RQ}) - N_i(\mathbf{R}) - N_i(\mathbf{Q}) \geq 0$ among all hydrophobic classes $\mu_i > 0$. Similarly, to make $\sum_{\mu_i<0} \mu_i [N_i(\mathbf{RQ}) - N_i(\mathbf{R}) - N_i(\mathbf{Q})]$ negative, there should be more $N_i(\mathbf{RQ}) - N_i(\mathbf{R}) - N_i(\mathbf{Q}) \geq 0$ than $N_i(\mathbf{RQ}) - N_i(\mathbf{R}) - N_i(\mathbf{Q}) < 0$ among all hydrophilic classes $\mu_i < 0$. Thus the best chance of $\Delta G_{ST} < 0$ is that $S \subset \Sigma_{\mathbf{R},h} = \bigcup_{\mu_i>0} \Sigma_{\mathbf{R},i}$, $T \subset \Sigma_{\mathbf{Q},h} = \bigcup_{\mu_i>0} \Sigma_{\mathbf{Q},i}$. Indeed, in that case, $N_i(\mathbf{RQ}) - N_i(\mathbf{R}) - N_i(\mathbf{Q}) = 0$ for each $\mu_i < 0$, hence $\sum_{\mu_i<0} \mu_i [N_i(\mathbf{RQ}) - N_i(\mathbf{R}) - N_i(\mathbf{Q})] = 0$; and $N_i(\mathbf{RQ}) - N_i(\mathbf{R}) - N_i(\mathbf{Q}) \leq 0$ for each $\mu_i > 0$; furthermore, since in that case it must be $\sum_{\mu_i>0} [N_i(\mathbf{RQ}) - N_i(\mathbf{R}) - N_i(\mathbf{Q})] = -N_r$, $\sum_{\mu_i>0} \mu_i [N_i(\mathbf{RQ}) - N_i(\mathbf{R}) - N_i(\mathbf{Q})] < 0$.

Let A_S (A_T) be the set of atoms in $\mathfrak{U}$ ($\mathfrak{V}$) such that these atoms (almost) touching S (T). Let $V_S \subset V_{\mathbf{R}}$ ($V_T \subset V_{\mathbf{Q}}$) be the subset corresponding to A_S (A_T). The potential difference $U(\mathbf{RQ}) - U(\mathbf{R}) - U(\mathbf{Q})$ is essentially the potential between V_S and V_T, $U(V_S, V_T)$. To make $U(V_S, V_T)$ negative or small positive, we want that S and T are complementary both in geometry and in elec-

trostatic distribution, such that the distances between atoms of A_S and A_T in the conformation $\mathbf{RQ}$ are small, and positive and negative charges can be matched to produce negative or small positive electrostatic potentials, and polar moieties may match to form hydrogen bonds to further reduce docking free energy.

In practice, we may find that connected pieces $S \subset \Sigma_{\mathbf{R}}$ and $T \subset \Sigma_{\mathbf{Q}}$ almost contained in $\Sigma_{\mathbf{R},h}$ and $\Sigma_{\mathbf{Q},h}$ except containing some tiny hydrophilic pieces (hydrophilic islands) at which a hydrophilic moiety has a hydrogen bond with water, or a charge neutralized by water. We will call such S and T as hydrophobic pieces with hydrophilic islands. So we should search the largest such hydrophobic pieces with hydrophilic islands as candidates of docking site S and T, if they are complementary both in geometry and in electrostatic distribution. Indeed, though hydrophilic islands means water molecules with $\mu_i < 0$ being removed, geometric complementary and matching hydrophilic islands in S and T to either neutralize charges or form hydrogen bonds can reduce $U(V_S, V_T)$ to compensate or at least partially compensate the increased $\sum_{\mu_i} \mu_i [N_i(\mathbf{RQ}) - N_i(\mathbf{R}) - N_i(\mathbf{Q})]$, and keep $\Delta G_{ST} < 0$.

3.4.4. Single Molecule Binding Affinity

If $\Delta G_{ST} < 0$, STH predicts that $\mathbf{RQ}$ will spontaneously fold to a stable conformation $\mathbf{R}_1\mathbf{Q}_1$ of the complex $\mathfrak{UW}$. $\mathbf{R}_1\mathbf{Q}_1$ is a (local or global) minimizer of $G(\mathbf{ML}; \mathfrak{UW}, \mathfrak{W})$, so $G(\mathbf{R}_1\mathbf{Q}_1; \mathfrak{UW}, \mathfrak{W}) < G(\mathbf{RQ}; \mathfrak{UW}, \mathfrak{W})$. Thus, we have

$$\Delta G_{\text{bind}} = G(\mathbf{R}_1\mathbf{Q}_1; \mathfrak{U}, \mathfrak{E}) - [G(\mathbf{R}; \mathfrak{U}, \mathfrak{E}) + G(\mathbf{Q}; \mathfrak{V}, \mathfrak{E})] < \Delta G_{ST} < 0. \tag{6}$$

The single molecule binding affinity is defined as

$$K = \exp\left(\frac{-\Delta_{\text{bind}}}{k_B T}\right),$$

where k_B is the Boltzmann constant and T the temperature in K˚. Because of $\Delta G_{\text{bind}} < 0$, $K > 1$, the bigger the K, the stronger the binding.

3.5. Global Geometric Featuers of Monomeric Globular Proteins

Based on the assumption (2), applying the formula in (4), we can predict that the native structure of a monomeric globular protein will have the following global geometric features:

1) It is compactly packed or it has a high packing density.

2) It has a globular shape, *i.e.*, looks like an ellipsoid even sphere and its surface area is small comparing with other conformations.

3) Hydrophobic moieties trend to hide from water as much as possible.

These global geometric features of monomeric globular protein native structures have been well known, only via Formula (4) the physical laws behind these global geometric features become clear, they are all the results of minimization of the CGF $G(\mathbf{R}, \Sigma_{\mathbf{R}}; \mathfrak{U}, \mathfrak{W})$. And, as will be shown, even nobody knows them, we can infer them from Formula (4).

3.5.1. Small Volume $V(\Omega_{\mathbf{R}})$ Means High Packing Density

Our first prediction of the global geometric feature of the native structure of a monomeric globular protein is that it must have a high packing density.

Let us clarify the relationship between density and volume by global geometry. Suppose that some quantity Q is distributed inside some domain Ω, *i.e.*, out of Ω the quantity Q is zero. Then density is a measure of quantity Q per volume unit expressed as $Q/V(\Omega)$.

Therefore, to measure density globally, the quantity Q and domain Ω are necessarily required. We have talked that a conformation $\mathbf{R}$ occupies a space $V_{\mathbf{R}}$, here the quantity Q is the volume $V(V_{\mathbf{R}})$. We also need a packing container to put $V_{\mathbf{R}}$ in. As stated in introduction, $V_{\mathbf{R}} \subset \bar{\Omega}_{\mathbf{R}} \subset \mathfrak{S}_{\mathbf{R}}$, thus, the container is $\bar{\Omega}_{\mathbf{R}}$.

Physically $V_{\mathbf{R}}$ is packed into the tailor made container $\bar{\Omega}_{\mathbf{R}}$, how effectively we packed the molecule $\mathfrak{U}$ then is measured by the packing density,

$$PD(\mathbf{R}) = \frac{V(V_{\mathbf{R}})}{V(\Omega_{\mathbf{R}})}, \tag{7}$$

Note that $V(V_{\mathbf{R}}) \le V(\bar{\Omega}_{\mathbf{R}}) = V(\Omega_{\mathbf{R}})$, and $V(V_{\mathbf{R}}) = V(\Omega_{\mathbf{R}})$ if and only if $\Sigma_{\mathbf{R}} = \partial V_{\mathbf{R}}$, thus, $PD(\mathbf{R})$ is always less than 1 unless $\Sigma_{\mathbf{R}} = \partial V_{\mathbf{R}}$. Because the van der Waals force, in nature a water molecular cannot really touching $\partial V_{\mathbf{R}}$, so the interface should not be selected as $\partial V_{\mathbf{R}}$. Thus, $PD(\mathbf{R})$ is always less than 1.

We will replace V_R by its good approximation $P_{\mathbf{R}}$ as explained in Method. Note that since $P_{\mathbf{R}} = \bigcup_{i=1}^{n} B(\mathbf{r}_i, r_i)$,

$$V(P_{\mathbf{R}}) = \sum_{i=1}^{n} V(B(\mathbf{r}_i, r_i)) - \text{overlapping volumes.}$$

For all conformations, $\sum_{i=1}^{n} V(B(\mathbf{r}_i, r_i))$ is a constant. By the steric condition (1), two balls $B(\mathbf{r}_i, r_i)$ and $B(\mathbf{r}_j, r_j)$ is overlapping, if and only if the atoms $\mathbf{a}_i$ and $\mathbf{a}_j$ are covalently bonded, *i.e.*, the distance $r_{ij} = |\mathbf{r}_i - \mathbf{r}_i| \ge r_i + r_j$ for not covalently bonded atoms $\mathbf{a}_i$ and $\mathbf{a}_j$, $V[B(\mathbf{r}_i, r_i) \cap B(\mathbf{r}_j, r_j)] = 0$. For covalently bonded $\mathbf{a}_i$ and $\mathbf{a}_j$, since $r_{ij} < r_i + r_j$ is always around the standard bond length as required in (1), r_{ij} will be almost a constant in all conformations, thus the overlapping volume $V[B(\mathbf{r}_i, r_i) \cap B(\mathbf{r}_j, r_j)]$ will almost be the same in all conformations. In rare cases that there are extra covalent bonds such as the disulfide bond in some conformations, the overlapping volume will increase a little. We conclude that the overlapping volume is almost the same for all conformations. Therefore, the volume $V(P_{\mathbf{R}})$ is almost the same for all conformations and equation (7) shows that the volume $V(\Omega_{\mathbf{R}})$ alone determines the packing density $PD(\mathbf{R})$. Therefore, shrinking $V(\Omega_{\mathbf{R}})$ is the only way to increase the packing density $PD(\mathbf{R})$.

Since $\omega_e > 0$, shrinking $V(\Omega_{\mathbf{R}})$ will reduce $G(\mathbf{R}, \Sigma_{\mathbf{R}}; \mathfrak{U}, \mathfrak{W})$ and simultaneously enlarge the packing density. Let $\mathbf{R}_{\mathfrak{U}}$ be the native structure of $\mathfrak{U}$, then by (2), for any other $\mathbf{R}$, $G(\mathbf{R}_{\mathfrak{U}}, \Sigma_{\mathbf{R}}; \mathfrak{U}, \mathfrak{W}) \le G(\mathbf{R}, \Sigma_{\mathbf{R}}; \mathfrak{U}, \mathfrak{W})$. Due to the effects of other terms in the Formula (4), $V(\Omega_{\mathbf{R}_{\mathfrak{U}}})$ may not be the minimum volume, but the packing density surely is as high as possible.

Indeed, high packing density was observed half century ago. In 1974, Richards observed that protein interiors are closely packed and he showed the interior of lysozyme and ribonuclease having a packing density of 0.75 ([20]). But the definition of packing density is not the global one given in (7), it was via complicated local calculations to determine each atom occupy how much space. Only much later the relations between packing density and volume was studied, [31].

The global definition of packing density (7) shows that shrinking $V(\Omega_R)$ must result the packing density becoming uniformly or homogenously high such that water molecules should not appear inside Ω_R since that will lower the packing density. This also was noted long ago, for example, it was observed that in general, water molecules are excluded from the interior of globular proteins, [20] [32] [33]. Unfilled cavities large enough to accommodate a water molecule only appear in very few cases, [20] [33]-[35]. In the literature, phrases such as "knobs in holes" and "ridges in grooves" are often used to describe the compactness (high packing density) of the packing, [32] [34] [36] [37], see **Figure 1**.

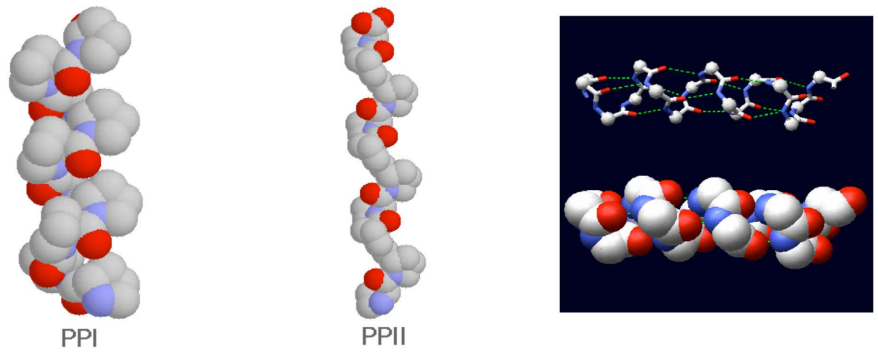

Figure 1. Left: Polyproline helices PPI (a right hand screw) and PPII (a left hand screw). Hydrogen bond is not necessary for a helix. Taking from [46]. Right: Taking from [47]. These screws make the packing very tight by occupying smaller space.

3.5.2. Small Surface Area Resulting Globular Shape

Our second prediction of the global geometric features of a monomeric globular protein's native structure is that it is shaped like a globule because that if volume $V(\Omega_R)$ is the conformation packer, the surface area $A(\Sigma_R)$ is the conformation shaper. This is because the **Isoperimetric Inequality:** Among all regions $\Omega \subset \mathbb{R}^3$ with finite volume $V = V(\Omega)$ and boundary area $A = A(\partial\Omega)$,

$$V^{\frac{2}{3}} \leq (36\pi)^{-\frac{1}{3}} A, \text{ equality holds if and only if } \Omega \text{ is a round ball.} \tag{8}$$

Alternative statements of the isopremetric inequality is that: 1) among all regions with the same volume $V_0 > 0$ and finite boundary surface area, only the round ball of radius $r = \left(\frac{3V_0}{4\pi}\right)^{\frac{1}{3}}$ has the smallest boundary surface area $A = 4\pi r^2$; 2) among all closed surfaces with the same surface area $A_0 > 0$, only the sphere with radius $r = \left(\frac{A_0}{4\pi}\right)^{\frac{1}{2}}$ bounds the largest volume $V = \frac{4}{3}\pi r^3$.

The isoperimetric inequality gives us a theoretical method to make a perfect sphere. Take a dough and change its shape (the volume V of the dough is preserved in changing), measure the boundary surface area A and calculate $C=\frac{V^{\frac{2}{3}}}{A}$, it is always $C\geq(36\pi)^{-\frac{1}{3}}$. If for some shape we have $C=(36\pi)^{-\frac{1}{3}}$, without looking at the dough, we know that we have got a perfect sphere by shaping the dough to a round ball.

Of course, in practice, nobody will make ball or sphere with such a clumsy procedure. But for protein structure, since $\omega_e>0$, by (4), shrinking $A(\Sigma_{\mathbf{R}})$ will reduce $G(\mathbf{R},\Sigma_{\mathbf{R}};\mathfrak{U},\mathfrak{W})$. Since $V(\Omega_{\mathbf{R}})\geq V(P_{\mathbf{R}})$ and $V(P_{\mathbf{R}})$ is almost a constant for any conformations, the isoperimetric inequality implies that to make $A(\Sigma_{\mathbf{R}})$ smaller and smaller, the $\Omega_{\mathbf{R}}$ must go towards to a round ball, otherwise, $A(\Sigma_{\mathbf{R}})$ would not keep shrinking as (2) required.

On the other hand, isoperimetric inequality also shows that $A(\Sigma_{\mathbf{R}})>(36\pi)^{\frac{1}{3}}[V(P_{\mathbf{R}})]^{\frac{2}{3}}$. Since $V(P_{\mathbf{R}})$ is almost a constant, one cannot make $A(\Sigma_{\mathbf{R}})$ too small.

Like the high packing density, it is observed long ago that the native structure of a globular protein has smaller (solvent accessible) surface area than that of other conformations ([19] [32] [33]). Janin ([38], an appendix to [39]) observed the need of "globular proteins to achieve a minimum accessible surface area compatible with their mass." During the folding of a fully extended polypeptide chain to give the compact native structure for proteins with a molecular weight of 15,000, the (solvent accessible surface) area decreases to about one-third of its maximum value. The $A(\Sigma_{\mathbf{R}})$ is shrunk so much, it is observed that even the accessible area of charged groups (hydrophilic surface area $A(\Sigma_{\mathbf{R},p})$ in our case) is substantially lower in the native protein (40% - 60% of $A(\Sigma_{\mathbf{R}})$) than in the extended chain, see [32]. This decreasing of $A(\Sigma_{\mathbf{R},p})$ will enlarge the $G(\mathbf{R},\Sigma_{\mathbf{R}};\mathfrak{U},\mathfrak{W})$ according to (4), one more example that the shrinking of $G(\mathbf{R},\Sigma_{\mathbf{R}};\mathfrak{U},\mathfrak{W})$ has to be cohesive. The $A(\Sigma_{\mathbf{R}})$ is decreased so much, thus $A(\Sigma_{\mathbf{R},p})$ has to be shrunk too, the decreased Gibbs free energy via decreasing $A(\Sigma_{\mathbf{R}})$ and $A(\Sigma_{\mathbf{R},h})$ seems more than enough to compensate the increased Gibbs free energy via shrinking $A(\Sigma_{\mathbf{R},p})$.

Novotny *et al.* [40] and [41] constructed incorrect conformations by putting the side chains from the N-end to the C-end of a globular protein (sea-worm hemerythrin) into the backbone of the native structure of another globular protein (variable domain of mouse immunoglobulin κ-chain) of the same length (113 residues), and vice versa, to get two incorrect conformations of proteins with known native structures. Comparing the incorrect conformations with native structures, they found that the solvent accessible surface areas of the incorrect conformations are significantly larger than that of the native structures. Many others also observed that smaller surface area phenomenon and by using various different models, concluded that it is a structural feature of the native

structures of globular proteins, see, for example, [33] [42]-[44].

3.5.3. Hydrophobic Moieties Tend to Avoid Contacting with Water

By Formula (4), besides shringking the intra-conformational potential $U(\mathbf{R})$ and the electron quantum terms $\omega_e V(V_{\mathbf{R}}) + d_w \omega_e A(\Sigma_{\mathbf{R}})$, shrinking the hydrophobic area $A(\Sigma_{\mathbf{R},h}) = \sum_{\mu_i>0} A(\Sigma_{\mathbf{R},i})$ and enlarging the hydrophilic area $A(\Sigma_{\mathbf{R},h}) = \sum_{\mu_i<0} A(\Sigma_{\mathbf{R},i})$ also contribute to reducing the Gibbs free energy $G(\mathbf{R}, \Sigma_{\mathbf{R}}; \mathfrak{U}, \mathfrak{W})$. Thus, by (2), besides small volume and surface area, we can predict that the native structure of a monomeric globular protein should have a small hydrophobic portion and large hydrophilic portion in its surface. Or, in other words, hydrophobic moieties tend to avoid contacting with water.

3.5.4. Hydrophobic Core

In fact, further mathematical argument not involving (4) can show that if the polypeptide chain length N is between 40 and 250, there will be hydrophobic core that is untouched by water molecules. For $N > 250$ the structure will be composed by domains connected by loops, each domain having a globular shape and tightly packed hydrophobic core, we will leave the long geometric argument to elsewhere. Here we only point out that

$$\text{Core}(\mathbf{R}) = \frac{A(\Sigma_{\mathbf{R},h})}{A(\Sigma_{\mathbf{R}})} \tag{9}$$

is an indicator of how well a hydrophobic core is formed, $\mathbf{Core}(\mathbf{R}) = 1$ means all water exposed moieties are hydrophobic, $\mathbf{Core}(\mathbf{R}) = 0$ means a perfect hydrophobic core, *i.e.*, not even one hydrophobic moiety is exposed to water.

3.6. Dominate Folding Force

3.6.1. Hyrophobic Force Comes From Aqueous Environment

In §3.5 we have shown that the native structure of a monomeric globular protein should be compactly packed and shaped like a globule with hydrophobic moiety hiding from water as much as possible. By formula of $G(\mathbf{R}; \Sigma_{\mathbf{R}}; \mathfrak{U}, \mathfrak{W})$ in (4), shrinking $A(\Sigma_{\mathbf{R},h})$ and enlarging $A(\Sigma_{\mathbf{R},p})$ have the same result, reducing the Gibbs free energy. Shrinking $A(\Sigma_{\mathbf{R},h})$ means that hydrophobic moieties having less chance to expose to water and indicating a hydrophobic effect. But enlarging $A(\Sigma_{\mathbf{R},p})$ means that hydrophilic moieties will have more chance to expose to water. Because that $A(\Sigma_{\mathbf{R}}) = A(\Sigma_{\mathbf{R},h}) + A(\Sigma_{\mathbf{R},p})$ and $A(\Sigma_{\mathbf{R}})$ is also shrinking to reduce the Gibbs free energy, enlarging $A(\Sigma_{\mathbf{R},p})$ will indirectly cause hydrophobic moieties having less chance to contact water, the same effect as the hydrophobic effect. Thus, Formula (4) shows the idea that there is a hydrophilic force or hydrophilic effect, [45], has a reason.

We can consider them together as the hydrophobic force, since they comes from the aqueous environment, we can also call them the aqueous force, or in general, environment force. In this sense, we may say that hydrophobic force is a force of protein folding. Is it the dominate folding force as claimed in [42]? It certainly is an important folding force.

3.6.2. Quantum Force

In §3.5 we have shown that the volume term $\omega_e V(\Omega_{\mathbf{R}})$ in (4) is the structure packer and the area term $d_w\omega_e A(\Sigma_{\mathbf{R}})$ is the structure shaper, without them we can hardly imaging the native structures of monomeric globular proteins are tightly packed and look like a glouble. Thus $\omega_e V(\Omega_{\mathbf{R}})$ and $d_w\omega_e A(\Sigma_{\mathbf{R}})$ are also major folding forces. As before mentioned, they come from quantum effect, let us call them quantum force.

3.6.3. Weak Forces

In ([25], pp. 110-114), there are four weak folding forces, or noncovalent bonds: "The folding of a protein chain is also determined by many different sets of weak noncovalent bonds that form between one part of the chain and another. These involve atoms in the polypeptide backbone, as well as atoms in the amino acid side chains. There are three types of these weak bonds: hydrogen bonds, electrostatic attractions, and van der Waals attractions, …" These weak forces are contained in $U(\mathbf{R})$ in formulas in (3) and (4). The fourth weak force in [25] is hydrophobic clustering force, just our hydrophobic force.

3.6.4. Dominate Folding Force

We claim that the quantum force and the aqueous force together is the dominate folding force. It shapes the native structure of a monomeric globular proteins such that it looks like a globule, it packed the native structure compactly and hiding the hdyrophobic moieties inside the interior as a hydrophobic core.

There are strong believes that hydrogen bonds should be dominate folding force. Hydrogen bonds essentially comes from electrostatic distribution, therefore, Intra-molecular hydrogen bonds, electrostatic attractions, and van der Waals attractions are countered in the intra-conformational potential $U(\mathbf{R})$. Intermolecular hydrogen bonds are counted in aqueous force.

Since intra or inter molecular bydrogen bonds reduce similar amount of energy, $U(\mathbf{R})$ cannot be the dominate folding force. But it is necessary in stabilizing the folded native structure due to the fact that proteins are special among polypeptide chains.

3.6.5. Proteins Are Special

Due to the existence of main chain hydrophilic moieties (the NH and CO groups), even $\mathbf{Core}(\mathbf{R}) = 0$, there are still hydrophilic portion inside the interior of the protein, the hydrophobic core is never pure hydrophobic. These main chain interior hydrophilic moieties have to form intramolecular hydrogen bonds, otherwise the intra-conformational potential $U(\mathbf{R})$ will be too large such that the conformation is not stable. Secondary structures such as helices, strands, and sheets are not only packed tightly, but also supply spacial arrangements of the entire polypeptide chain that are able to form regular hydrogen bonding patterns associated with secondary structures. Besides the main chain hydrophilic moieties, if there are hydrophilic side chains in the interior of the protein, they have to be arranged in space such that all polar and charged interior side chains

would be nearby enough to each other to neutralize electron charges, or to form hydrogen bonds, to keep $U(\mathbf{R})$ low. The longer the polypeptide chain, the more difficult to arrange such a delicate and balanced spacial arrangement of the entire polypeptide chain.

Because of this delicate and balanced spacial arrangement of the polypeptide chain for native structures of monomeric globular proteins, the polypeptide chains of them must be very special, and perhaps is selected by natural selection for just this advantage.

Not every polypeptide chain can have this kind of delicate and balanced spacial arrangement, *i.e.*, polypeptide chains of proteins are very special, they are selected by nature selection. It is estimated in [6] that for a random polypeptide chain $\mathbf{A}$ of 400 residues, the probability $P(\mathbf{A})$ of $\mathbf{A}$ is a protein's polypeptide chain is

$$P(\mathbf{A}) \le 10^{60} / 10^{520} = 10^{-460},$$

$P(\mathbf{A})$ is practically zero. The monomeric globular proteins are even more special.

3.7. A Wholistic View of Native Structures of Monomeric Globular Proteins

By assumption (2), the global minimizer of $G(\mathbf{R}, \Sigma_{\mathbf{R}}; \mathfrak{U}, \mathfrak{W})$ is the native structure $\mathbf{R}_{\mathfrak{U}}$. The dominate folding force, quantum force and aqueous force, are global ones, *i.e.*, only taking the protein molecule as a whole we can measure them. Shrinking or enlarging global measures such as volume and area and hydrophobic and hydrophilic areas cohesively (in synergy) is a global folding force. As shown above, under these global forces the entire polypeptide chain has to be arranged in a delicate and balanced spacial arrangement, only a vanishingly tiny portion of polypeptide chains can be polypeptide chains of monomeric globular proteins.

As discussed above, the delicate and balanced spacial arrangement of the entire polypeptide chain prefers secondary structures such as α helices and β strands, and semi-local structures such as the pleated β sheets, with their regular hydrogen bond patterns.

Because of these global forces, we suggest a point of view of native structures of monomeric globular proteins, a wholistic top-down point of view: The entire molecule as a whole determines the local structures, *i.e.*, global folding forces produce a compactly packed tertiary structure that looks like a globule. Secondary structures appear as by-products of this global folding force pushing high packing density and keep $U(\mathbf{R})$ as low as possible via the speciality of monomeric globular protein polypeptide chains that can be arranged a delicate and balanced spacial arrangement. Moreover, under this global folding force, this delicate and balanced spacial arrangement also hides hydrophobic moieties away from water as much as possible.

On the contrary, a bottom-up point of view of native structures of proteins

believes that secondary structures appear due to the propensities of short peptide chains independent of the whole polypeptide chain, and the tertiary structure is formed by a suitable arrangement of these secondary structures connected by turns and loops.

In fact, a wholistic point of view was already implied in Anfinsen's Thermodynamic Hypothesis, "the native conformation is determined by the **totality of interatomic interactions** and hence by the amino acid sequence, in a given environment." [3].

Allosterey can make long distance conformational change, and other facts in the following subsections, PPI and PPII helices without hydrogen bonding, simulations by reducing hydrophobic surface alone can produce secondary structures and hydrogen bonds, chameleon sequences, point mutation causes dramatical conformation change, all of them support the wholistic top-down point of view. The bottom-up point of view cannot explain these phenomena.

3.7.1. Secondary Structures Are Due to Compact Packing

Regular intramolecular hydrogen bond patterns associated with α helix and pleated β sheet lowered the intra-conformational potential $U(\mathbf{R})$ in (3) and (4), thus $G(\mathbf{R};\mathfrak{U},\mathfrak{W})$ or $G(\mathbf{R},\Sigma_{\mathbf{R}};\mathfrak{U},\mathfrak{W})$ is lowered. These intramolecular hydrogen bond patterns give a reason to support of the bottom-up point of view of protein structures.

The wholistic top-down point of view suggests that it is the compact packing that produces the secondary structures, hydrogen bond patterns associated with secondary structures are only by-products of this compact packing via a delicate and balanced spacial arrangement of the entire polypeptide chain, and only a vanishing portion of polypeptide chains can afford this delicate and balanced spacial arrangement. While screws (helices) and strands being formed to get compact packing, regular hydrogen bond patterns can be simultaneously realized due to this delitcate and balanced spacial arrangement of these naturally selected special polypeptide chains.

From the geometric point of view, there are only very few methods to compactly pack a uneven rope (the extended, long and thin wire conformation of the polypeptide chain), one of them is to form some kind of screw shape, *i.e.*, a helix, another is to form straight strands and arrange these strand as sheets; these correspond to secondary structures. Turns and loops are necessary connecting portions of the wire between consecutive secondary structures, see [42]. Compactly pack a random polypeptide chain, we will find a lot of unmatched hydrogen bond donors and accepters, as well as unmatched charged side chains, leading to a large intra-confomational potential $U(\mathbf{R})$, therefore, a large Gibbs free energy. Due to unable to arrange a delicate and balanced spacial arrangement of the random polypeptide, the only way of lowering CGF is make all those hydrogen bond donors and acceptors to form hydrogen bonds with water molecules, as well as to neutralize charges with water. This will increase surface areas $A(\Sigma_{\mathbf{R}})$, $A(\Sigma_{\mathbf{R},h})$, and $A(\Sigma_{\mathbf{R},p})$, but the lowered $U(\mathbf{R})$ will be more than enough to

compensate. That is why these random polypeptide chains are not proteins. Instead of natives structures, they will have many extended stable conformations, just like water soluble polymers. On the other hand, without main chain hydrophilic moieties, a polyproline peptide in a protein can form helices such as PPI and PPII as shown in **Figure 1**. Of course, these PPI and PPII are formed by compact packing and do not have any hydrogen bond.

The common geometric characteristic of PPI and PPII, as well as the regular α helices, is that their screw pattern makes the structure denser, or packed more tightly to reduce the occupied spaces to as small as possible. Once the concept of helices be modified accordingly, very simplified models such as the HP model [42] and tube model [48] obtained "secondary structures" like helices and sheets, of course no hydrogen bonds exist because of the simplicity of these models, but indeed the packing density is high. Thus, intramolecular hydrogen bond patterns are not necessary of secondary structures, instead, they are by-products of a global packing force to the specially selected protein polypeptide chains.

3.7.2. An *ab initio* Simulation of Shrinking Hydrophobic Area Alone Produces Secondary Structures and Hydrogen Bonds

An *ab initio* simulation of shrinking the hydrophobic surface area $A(\Sigma_{\mathbf{R},h})$ alone has produced secondary structures such as α helices and β strands and their associated intramolecular hydrogen bonds, with statistical significance, [49]. This *ab initio* simulation shows that a global force alone can really produce secondary structures, supporting the wholistic top-down point of view of monomeric globular protein structures.

It is admitted in [50], hydrogen bonding must be explicitly modelled for helix formation, otherwise molecular dynamics simulation based on pairwise potential energy cannot produce secondary structures.

3.7.3. Chameleon Sequences

A chameleon sequence is a short peptide sequence that in native structures of different proteins may appear as different secondary structurss, for example, α helix in protein $\mathfrak{U}$, and β strand in protein $\mathfrak{V}$.

Note that in the discussion of mechanism of allostery above, although $\mathbf{R}_1\mathbf{Q}_1$ is a stable conformation of $\mathfrak{UV}$, but $\mathbf{R}_1$ and $\mathbf{Q}_1$ may not be a stable conformation of $\mathfrak{U}$ and $\mathfrak{V}$. Global ($\mathfrak{UV}$) stability does not guarantee local ($\mathfrak{U}$ and $\mathfrak{V}$) stabilities. Even though $\mathbf{R}_1$ and $\mathbf{Q}_1$ are not stable as parts of $\mathbf{R}_1\mathbf{Q}_1$, the compensation is that $\mathbf{R}_1\mathbf{Q}_1$ is stable. On the other hand, $\mathbf{R} \in \mathfrak{X}_{\mathfrak{U}}$ and $\mathbf{Q} \in \mathfrak{X}_{\mathfrak{V}}$ are both stable, together $\mathbf{RQ} \in \mathfrak{X}_{\mathfrak{UV}}$ is not stable. Thus, local stabilities do not guarantee global stability either.

These observations suggest that local conformations has to adjust to one another in space to achieve global stability, therefore, suggest that the folding force must be a global one, the entire polypeptide chain(s) must be considered together to achieve a stable conformation. Indeed, in the water environment $\mathfrak{W}$, look at the CGF Formula (4), volume and area, and hydrophobic and hydrophilic areas, all are truly global, one cannot define them without onsidering

the entire polypeptide chain.

A chameleon sequence has to take a α helix secondary structure to help achieving a stable conformation $\mathbf{R}_{\mathfrak{U}}$ of $\mathfrak{U}$; but the same chameleon sequence has to take a β strand secondary structure to help to achieve a stable $\mathbf{R}_{\mathfrak{V}}$ of $\mathfrak{V}$, even thought the β strand may less stable than α helix, the compensation is that $\mathbf{R}_{\mathfrak{V}}$ is stable. Stability of the conformation as a whoe requires the same chameleon sequence taking different secondary structures in different proteins.

3.7.4. Point Mutation

Point mutation, or single mutation, is the substitution of just one residue of a polypeptide chain. Although just one residue is substituted, point mutation may cause dramatic conformational change. The reference [51] is entitled a "One sequence plus one mutation equals to two folds." The reference [52] is entitled as "Structure of the Hydrophobic Core Determines the 3D Protein Structure-Verification by Single Mutation Proteins", its abstract states: "Four de novo proteins differing in single mutation positions, with a chain length of 56 amino acids, represent diverse 3D structures: monomeric 3α and $4\beta+\alpha$ folds. The reason for this diversity is seen in the different structure of the hydrophobic core as a result of synergy leading to the generation of a system in which the polypeptide chain as a whole participates."

4. Conclusions

We have demonstrated that CGF and STH can make predictions and explanations, but only qualitatively. Formulas of CGF in (3) and (4) have opened the door to quantitative investigations to resolve old problems in life science, such as the protein foliding problem, *ab initio* prediction of the native structure of a water soluble protein. Quantitative calculating the conformational changes in allostery is essentially also a folding problem. Resolutions of these problems will make not only theoretical progresses but also will have many practical applications. But, a obstacle remains.

Looking at Formula (4), because $\omega_e > 0$, one way to shrink $G(\mathbf{R},\Sigma_{\mathbf{R}};\mathfrak{U},\mathfrak{E})$ is to change $\mathbf{R}$ such that the volume $V(\Omega_{\mathbf{R}})$ and surface area $A(\Sigma_{\mathbf{R}})$ decrease. Similarly, we want the hydrophobic surface area $A(\Sigma_{\mathbf{R},h}) = A(\bigcup_{\omega_i>0}\Sigma_{\mathbf{R},i}) = \sum_{\omega_i>0} A(\Sigma_{\mathbf{R},i})$ to get as small as possible and the hydrophilic surface area $A(\Sigma_{\mathbf{R},p}) = A(\bigcup_{\omega_i<0}\Sigma_{\mathbf{R},i}) = \sum_{\omega_i<0} A(\Sigma_{\mathbf{R},i})$ to get as large as possible, to shrink $G(\mathbf{R},\Sigma_{\mathbf{R}};\mathfrak{U},\mathfrak{E})$. There are some restrictions to these decreases and increases, for example, since $A(\Sigma_{\mathbf{R},p}) \le A(\Sigma_{\mathbf{R}})$, $A(\Sigma_{\mathbf{R},p})$ cannot become too large while $A(\Sigma_{\mathbf{R}})$ is shrinking with changing conformations. Indeed, to make $G(\mathbf{R},\Sigma_{\mathbf{R}};\mathfrak{U},\mathfrak{W})$ as small as possible, we have to balance between volume, area, hydrophobic area, and hydrophilic area; as well as the intra-conformational potential $U(\mathbf{R})$. Thus, we should cohesively simultaneously shrinking the volume $V(\Omega_{\mathbf{R}})$, the area $A(\Sigma_{\mathbf{R}})$, and the hydrophobic surface area $A(\Sigma_{\mathbf{R},h})$; as well as to shrink the intra-conformational potential $U(\mathbf{R})$ to

make more intramolecular hydrogen bonds, neutralised charges, and to avoid overlapping of no covalent bonded atoms. If $A(\Sigma_{\mathbf{R}})$ and $A(\Sigma_{\mathbf{R},h})$ both become smaller, $A(\Sigma_{\mathbf{R},p}) = A(\Sigma_{\mathbf{R}}) - A(\Sigma_{\mathbf{R},h})$ will increasing or decreasing depending on the degrees of shrinking of $A(\Sigma_{\mathbf{R}})$ and $A(\Sigma_{\mathbf{R},h})$.

How to cohesively (in synergy) do all of these? We need the accurate values of μ_e, ω_e, μ_i, ω_i, $i = 1, \cdots, L$. Once we know these accurate values, one can readily do *ab initio* structure prediction of a globular protein via Formula (3) or (4). Thus, the *ab initio* structure predicting problem for globular proteins and accurately predict post-binding deformations will become pure mathematical minimization problems.

Unfortunately, so far, the obstacle is that there is no accurate determination of these chemical potentials. Determining these chemical potentials is the next task towards resolving these famous problems by first principle. The determination methods will be both theoretical and experimental. Heavy computations in training on a set of native structures of monomeric globular proteins is necessary.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Wu, H. (1929) A Theory of Denatured and Coagulated Proteins. *America Journal of Physiology*, **90**, 562-563.

[2] Wu, H. (1931) Studies on Denaturation of Proteins XIII. A Theory of Denaturation. *Chinese Journal of Physiology*, **5**, 321-344.

[3] Anfinsen, C.B. (1973) Principles That Govern the Folding of Protein Chains. *Science*, **181**, 223-230. https://doi.org/10.1126/science.181.4096.223

[4] Ho, B., Baryshnikova, A. and Brown, G.W. (2018) Unification of Protein Abundance Datasets Yields a Quantitative *Saccharomyces cerevisiae* Proteome. *Cell Systems*, **6**, 192-205.e3. https://doi.org/10.1016/j.cels.2017.12.004

[5] Serdyuk, I.N., Zaccai, N.R. and Zaccai, J. (2007). Methods in Molecular Biophysics: Structure, Dynamics, Function. Cambridge University Press. https://doi.org/10.1017/cbo9780511811166

[6] Fang, Y. (2012) Gibbs Free Energy Formula for Protein Folding. In: Morales-Rodriguez, R., Ed., *Thermodynamics—Fundamentals and Its Application in Science*, IntechOpen, London, 47-82. http://www.intechopen.com/books/thermodynamics-fundamentals-and-its-application-in-science

[7] Fang, Y. (2013) Ben-Naim's "Pitfall": Don Quixote's Windmill. *Open Journal of Biophysics*, **3**, 13-21. https://doi.org/10.4236/ojbiphy.2013.31002

[8] Fang, Y. (2014) The Second Law, Gibbs Free Energy, Geometry, and Protein Folding. *Journal of Advances in Physics*, **3**, 278-285. https://doi.org/10.24297/jap.v3i3.2061

[9] Fang, Y. (2014) A Gibbs Free Energy Formula for Protein Folding Derived from

Quantum Statistics. *Science China Physics, Mechanics & Astronomy*, **57**, 1547-1551. https://doi.org/10.1007/s11433-013-5288-x

[10] Fang, Y. (2015) Thermodynamic Principle Revisited: Theory of Protein Folding. *Advances in Bioscience and Biotechnology*, **6**, 37-48. https://doi.org/10.4236/abb.2015.61005

[11] Fang, Y. (2015) Why Ben-Naim's Deepest Pitfall Does Not Exist. *Open Journal of Biophysics*, **5**, 45-49. https://doi.org/10.4236/ojbiphy.2015.52004

[12] Fang, Y. (2015) Protein Folding: Towards a Single Molecule Theory. In: Hale, M., Ed., *Advances in Protein Folding Research*, Nova Publishers, New York, 65-126.

[13] Yi, F. (2019) Single Molecule Thermodynamics Hypothesis of Protein Folding and Drug Design. *Journal of Biosciences and Medicines*, **7**, 164-172. https://doi.org/10.4236/jbm.2019.711015

[14] Schafer, N.P., Kim, B.L., Zheng, W. and Wolynes, P.G. (2014) Learning to Fold Proteins Using Energy Landscape Theory. *Israel Journal of Chemistry*, **54**, 1311-1337. https://doi.org/10.1002/ijch.201300145

[15] Ben-Naim, A. (2011) Pitfalls in Anfinsen's Thermodynamic Hypothesis. *Chemical Physics Letters*, **511**, 126-128. https://doi.org/10.1016/j.cplett.2011.05.049

[16] Lazaridis, T. and Karplus, M. (1999) Effective Energy Function for Proteins in Solution. *Proteins: Structure, Function, and Genetics*, **35**, 133-152. https://doi.org/10.1002/(sici)1097-0134(19990501)35:2<133::aid-prot1>3.0.co;2-n

[17] Lazaridis, T. and Karplus, M. (2002) Thermodynamics of Protein Folding: A Microscopic View. *Biophysical Chemistry*, **100**, 367-395. https://doi.org/10.1016/s0301-4622(02)00293-4

[18] Bader, R.F.W. (1994) Atoms in Molecules: A Quantum Theory. Clarendon Press.

[19] Lee, B. and Richards, F.M. (1971) The Interpretation of Protein Structures: Estimation of Static Accessibility. *Journal of Molecular Biology*, **55**, 379-IN4. https://doi.org/10.1016/0022-2836(71)90324-x

[20] Richards, F.M. (1974) The Interpretation of Protein Structures: Total Volume, Group Volume Distributions and Packing Density. *Journal of Molecular Biology*, **82**, 1-14. https://doi.org/10.1016/0022-2836(74)90570-1

[21] Connolly, M.L. (1983) Analytical Molecular Surface Calculation. *Journal of Applied Crystallography*, **16**, 548-558. https://doi.org/10.1107/s0021889883010985

[22] Connolly, M.L. (2012) Molecular Surfaces. http://www.biohedron.com/

[23] Kessel, A. and Ben-Tal, N. (2018) Introduction to Proteins: Structure, Function, and Motion Second Edition. CRC Press.

[24] Fenton, A.W. (2008) Allostery: An Illustrated Definition for the "Second Secret of Life". *Trends in Biochemical Sciences*, **33**, 420-425. https://doi.org/10.1016/j.tibs.2008.05.009

[25] Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K. and Walter, P. (2015) Molecular Biology of the Cell. 6th Edition, Garland Science.

[26] Bloomberg, C. (2007) Physics of Life: The Physicist's Road to Biology. Elsevier.

[27] Thom, R. (1975) Structure Stability and Morphogenesis: An Outline of a General Theory of Models. W.A. Benjaming, Inc.

[28] Sanns, W. (2009) Catastrophe Theory. In: Meyers, R.A., Ed., *Encyclopedia of Complexity and System Science*, Vol. 4, Springer, 703-719.

[29] Fersht, A.R. (2008) From the First Protein Structures to Our Current Knowledge of Protein Folding: Delights and Scepticisms. *Nature Reviews Molecular Cell Biology*,

9, 650-654. https://doi.org/10.1038/nrm2446

[30] Uversky, V.N. (2019) Intrinsically Disordered Proteins and Their "Mysterious" (Meta)physics. *Frontiers in Physics*, **7**, Article No. 10. https://doi.org/10.3389/fphy.2019.00010

[31] Tsai, J., Taylor, R., Chothia, C. and Gerstein, M. (1999) The Packing Density in Proteins: Standard Radii and Volumes. *Journal of Molecular Biology*, **290**, 253-266.

[32] Richards, F.M. (1977) Areas, Volumes, Packing, and Protein Structure. *Annual Review of Biophysics and Bioengineering*, **6**, 151-176. https://doi.org/10.1146/annurev.bb.06.060177.001055

[33] Richards, F.M. (1979) Packing Defects, Cavities, Volume Fluctuations, and Access to the Interior of Proteins. Including Some General Comments on Surface Area and Protein Structure. *Carlsberg Research Communications*, **44**, 47-63. https://doi.org/10.1007/bf02906521

[34] Creighton, T.E. (1993) Proteins: Structures and Molecular Properties. 2nd Edition, W. H. Freeman and Company.

[35] Richards, F.M. and Lim, W.A. (1993) An Analysis of Packing in the Protein Folding Problem. *Quarterly Reviews of Biophysics*, **26**, 423-498. https://doi.org/10.1017/s0033583500002845

[36] Berg, J.M., Tymoczko, J.L. and Stryer, L. (2002) Biochemistry. 5th Edition, W.H. Freeman and Company.

[37] Finkelstein, A.V. and Ptitsyn, O.B. (2016) Protein Physics: A Course of Lectures. Second, Updated and Extended Edition, Academic Press.

[38] Janin, J. (1976) Surface Area of Globular Proteins. *Journal of Molecular Biology*, **105**, 13-14. https://doi.org/10.1016/0022-2836(76)90192-3

[39] Chothia, C. (1976) The Nature of the Accessible and Buried Surfaces in Proteins. *Journal of Molecular Biology*, **105**, 1-12. https://doi.org/10.1016/0022-2836(76)90191-1

[40] Novotný, J., Bruccoleri, R. and Karplus, M. (1984) An Analysis of Incorrectly Folded Protein Models. Implications for Structure Predictions. *Journal of Molecular Biology*, **177**, 787-818. https://doi.org/10.1016/0022-2836(84)90049-4

[41] Novotný, J., Rashin, A.A. and Bruccoleri, R.E. (1988) Criteria That Discriminate between Native Proteins and Incorrectly Folded Models. *Proteins*: *Structure*, *Function*, *and Bioinformatics*, **4**, 19-30. https://doi.org/10.1002/prot.340040105

[42] Dill, K.A. (1990) Dominant Forces in Protein Folding. *Biochemistry*, **29**, 7133-7155. https://doi.org/10.1021/bi00483a001

[43] Chan, H.S. and Dill, K.A. (1990) The Effects of Internal Constraints on the Configurations of Chain Molecules. *The Journal of Chemical Physics*, **92**, 3118-3135. https://doi.org/10.1063/1.458605

[44] Naderi-Manesh, H., Sadeghi, M., Arab, S. and Moosavi Movahedi, A.A. (2001) Prediction of Protein Surface Accessibility with Information Theory. *Proteins*: *Structure*, *Function*, *and Bioinformatics*, **42**, 452-459. https://doi.org/10.1002/1097-0134(20010301)42:4<452::aid-prot40>3.0.co;2-q

[45] Ben-Naim, A. (2011) The Rise and Fall of the Hydrophobic Effect in Protein Folding and Protein-Protein Association, and Molecular Recognition. *Open Journal of Biophysics*, **1**, 1-7. https://doi.org/10.4236/ojbiphy.2011.11001

[46] Description of Polyproline Helices. http://www.cryst.bbk.ac.uk/pps97/assignments/projects/szabo/pphelix.htm

[47] ExPASy Proteomics Server. http://au.expasy.org/sprot/relnotes/relstat.html

[48] Maritan, A., Micheletti, C., Trovato, A. and Banavar, J.R. (2000) Optimal Shapes of Compact Strings. *Nature*, **406**, 287-290. https://doi.org/10.1038/35018538

[49] Fang, Y. and Jing, J. (2010) Geometry, Thermodynamics, and Protein. *Journal of Theoretical Biology*, **262**, 383-390. https://doi.org/10.1016/j.jtbi.2009.09.013

[50] Hubner, I.A. and Shakhnovich, E.I. (2005) Geometric and Physical Considerations for Realistic Protein Models. *Physical Review E*, **72**, Article ID: 022901. https://doi.org/10.1103/physreve.72.022901

[51] Shortle, D. (2009) One Sequence plus One Mutation Equals Two Folds. *Proceedings of the National Academy of Sciences*, **106**, 21011-21012. https://doi.org/10.1073/pnas.0912370107

[52] Banach, M., Fabian, P., Stapor, K., Konieczny, L. and Roterman, A.I. (2020) Structure of the Hydrophobic Core Determines the 3D Protein Structure—Verification by Single Mutation Proteins. *Biomolecules*, **10**, Article No. 767. https://doi.org/10.3390/biom10050767