Protein structure along the order-disorder continuum

Charles K. Fisher[†] and Collin M. Stultz*[,†,‡]
Committee on Higher Degrees in Biophysics, Harvard University, Cambridge, MA 02139-4307
Harvard-MIT Division of Health Sciences and Technology, Department of Electrical Engineering and Computer Science and the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139-4307
E-mail: cmstultz@mit.edu
†Harvard University
‡Massachusetts Institute of Technology

Supplementary Text

Derivation of the IOP for protein ensembles

The ideal order parameter, *O*, for protein ensembles should have the following properties:

1. $0 \leq O \leq 1$
2. $O = 1$ iff the protein adopts one conformation throughout its biological lifetime;
3. $O = 0$ iff the protein can only be described by an infinite number of structurally dissimilar states; i.e. the protein is completely unstructured.

We begin by defining an ensemble by $X = (\mathbf{w}, S)$, where $S = \{s_1, \ldots, s_n\}$ is the set of conformations $s_i \in R^{3N}$ and $\mathbf{w} = (w_1, \ldots, w_n)$ is the corresponding vector of population weights, which is a function of the relative stabilities of the different structures within the ensemble. We begin by grouping these conformations into two "states": one consisting of a single conformation, s_i , and the other consisting of all the other conformations, $\{S_j\}_{j \neq i}$. The probability of the first state is w_i and the probability of the second state is $1 - w_i$. In fact, there are *n* different ways of partitioning the structures in the ensemble in this fashion. Our motivation comes from the realization that if the protein is completely ordered (which we will call "frozen") then one of these two states will have probability 1. That is, if the protein is frozen then there exists a k such that $w_k = 1$. We now compute the ensemble average distance between the

probability vectors $\vec{W}_i = (w_i, 1-w_i)$ and $\vec{F} = (1,0)$, where the latter corresponds to the protein being frozen:

$$\tilde{O} = 1 - \sum_{i=1}^{n} w_i D(\vec{W}_i, \vec{F}) \tag{1}$$

Here D is some appropriate measure of the distance between the probability vectors $\vec{W}_i$ and $\vec{F}$. Note that if the protein only adopts structure s_k then $w_k = 1$, $\vec{W}_k = (1,0)$ and $\forall_{i \neq k} \left(w_i = 0 \text{ and } \vec{W}_i = (0,1) \right)$ so that $\tilde{O} = 1$.

To implement the scheme outlined in equation (1), we need an appropriate choice of a distance metric between two probability distributions/vectors. The distance between two generic probability vectors, $\vec{P} = (p_1, \ldots, p_n)$ and $\vec{Q} = (q_1, \ldots, q_n)$ is typically obtained using the Kullback-Leibler (KL) divergence:[1]

$$KL\left(\vec{P} \middle\| \vec{Q}\right) = \sum_{i=1}^{m} p_i \log_2 \frac{p_i}{q_i} \tag{2}$$

which has a range $0 \leq KL\left(\vec{P} \middle\| \vec{Q}\right) < \infty$. However, as stated earlier, we need our order parameter to be bounded by 0 and 1. A slightly modified form of the KL divergence, $KL\left(\vec{P} \middle\| \frac{\vec{P}+\vec{Q}}{2}\right)$, which itself lies in the interval $[0,1]$, makes this possible. Now if we define

$$D(\vec{W}_i, \vec{F}) \equiv KL\left(\vec{F}_i \middle\| \frac{\vec{W}_i + \vec{F}}{2}\right) \tag{3}$$

then equation (1) becomes:

$$\tilde{O} \equiv 1 - \sum_{i=1}^{n} w_i KL\left(\vec{F}_i \middle\| \frac{\vec{W}_i + \vec{F}}{2} \right) = \sum_{i=1}^{n} w_i \log_2\left(1 + w_i\right) \qquad (4)$$

Formal definition of $\tilde{O}$ and its bounds

In the derivations presented above and throughout the main text we have focused on using the order parameter to describe the heterogeneity of a protein conformational ensemble; however, the OP can be defined more generally as a description of the degree of order for any discrete random variable.

Definition: Let *X* be a discrete random variable and *p* be a probability distribution of *X*. The information order parameter is defined as $\tilde{O}(p) = \sum_{x \in X} p(x) \log_2\left(1 + p(x)\right)$, where we have explicitly written the order parameter as a function of the underlying probability distribution.

Theorem 1: Let *n* be the number of different states available to the discrete random variable *X* with probability distribution *p*. The information order parameter is bounded by $\log_2 \frac{n+1}{n} \le \tilde{O}(p) \le 1$.

Proof of Theorem 1: First, note that since probabilities are non-negative $p(x) \ge 0$ and $\log_2\left(1 + p(x)\right) \ge 0$ implying that $\tilde{O}(p) \ge 0$. To prove the tighter lower bound we will minimize $\tilde{O}(p)$ over p subject to the constraint that $\sum_{x \in X} p(x) = 1$ using the method of Lagrange multipliers. The Lagrange function to be minimized is:

$$\Lambda(p,\lambda)=\sum_{x\in X}p(x)\log_2\left(1+p(x)\right)+\lambda\left(\sum_{x\in X}p(x)-1\right)$$

Taking the derivatives, we have $\log\left(1+p(x)\right)+\frac{p(x)}{1+p(x)}+\lambda=0$, which implies that $p(x)=p(y)\,\forall x,y\in X$ because the probability of $x\in X$ depends only on the Lagrange multiplier. From this, and the given summation constraint, we have $p(x)=\frac{1}{n}\,\forall x\in X$, which corresponds to $\tilde{O}(p)=\log_2\frac{n+1}{n}$. The upper bound follows from Jensen's inequality for concave functions as $\tilde{O}(p)=\sum_{x\in X}p(x)\log_2\left(1+p(x)\right)\le\log_2\left(1+\sum_{x\in X}p(x)^2\right)\le 1$ ■

A smoothed order parameter

The metric given by equation (4) has the correct bounds for an order parameter; however, it ignores the fact that structures within a given ensemble may share some similarity. If many of the structures happen to be very similar to each other than it is helpful to smooth the $\vec{W}_i$ probabilities with a kernel function. A simple method is to use a Gaussian kernel, giving a smoothed probability $\Omega(s_i)\equiv\sum_{j=1}^{n}w_j\exp\left(\frac{-D^2(s_i,s_j)}{2\langle D^2\rangle}\right)$, where $D^2(s_i,s_j)$ is the Cα coordinate mean-square deviation (MSD) between structures s_i and s_j and $\langle D^2\rangle$ is the average pairwise MSD due to the fluctuations of a typical protein, which we estimate from simulations as described below. Plugging the smoothed probabilities into eq. (4) we obtain:

$$O \equiv \sum_{i=1}^{n} w_i \log_2\left(1+\Omega\left(s_i\right)\right) \tag{5}$$

It is easy to show that both O and $\tilde{O}$ lie within $[0,1]$. Note that $0 \le w_i$, and $\Omega\left(s_i\right) \le 1$ implying that $0 \le \log_2\left(1+w_i\right)$ and $\log_2\left(1+\Omega\left(s_i\right)\right) \le 1$. Thus, $0 \le O$ and $\tilde{O} \le 1$ follows automatically from eqs. (4) and (5). In addition, we present a theorem below showing that $O \ge \tilde{O}$.

Theorem 2: Let $X = \left(\mathbf{w}, S\right)$ be an ensemble where $\mathbf{w}$ is a vector of population weights and $S = \left\{s_1, \ldots, s_n\right\}$ is a set of non-redundant structures; i.e. $D^2\left(s_i, s_j\right) = 0 \Leftrightarrow i = j$. Then $O \ge \tilde{O}$ for any value of $\left\langle D^2 \right\rangle > 0$.

Before presenting the proof of Theorem 2, it is necessary to state and prove the following simple lemmas.

Lemma 1: $\lim_{\left\langle D^2 \right\rangle \to 0} O = \tilde{O}$.

$$\begin{aligned}
\lim_{\left\langle D^2 \right\rangle \to 0} O &= \lim_{\left\langle D^2 \right\rangle \to 0} \sum_{i=1}^{n} w_i \log_2\left(1+\sum_{j=1}^{n} w_j \exp\left(\frac{-D^2\left(s_i, s_j\right)}{2\left\langle D^2 \right\rangle}\right)\right) \\
&= \sum_{i=1}^{n} w_i \log_2\left(1+\sum_{j=1}^{n} w_j \delta_{ij}\right) \\
&= \sum_{i=1}^{n} w_i \log_2\left(1+w_i\right) \\
&= \tilde{O}
\end{aligned}$$

Lemma 2: $\dfrac{\partial O}{\partial\left\langle D^2 \right\rangle} \ge 0$.

$$\frac{\partial O}{\partial \left\langle D^2 \right\rangle} = \sum_{i=1}^{n} w_i \frac{\partial}{\partial \left\langle D^2 \right\rangle} \log_2 \left(1 + \sum_{j=1}^{n} w_j \exp\left(\frac{-D^2\left(s_i, s_j\right)}{2\left\langle D^2 \right\rangle} \right) \right)$$

$$= \left(\log_2 e\right) \sum_{i=1}^{n} w_i \frac{\sum_{j=1}^{n} w_j \frac{D^2\left(s_i, s_j\right)}{2\left\langle D^2 \right\rangle^{-2}} \exp\left(\frac{-D^2\left(s_i, s_j\right)}{2\left\langle D^2 \right\rangle} \right)}{1 + \sum_{j=1}^{n} w_j \exp\left(\frac{-D^2\left(s_i, s_j\right)}{2\left\langle D^2 \right\rangle} \right)}$$

$$\geq 0$$

Proof of Theorem 2: Suppose that Theorem 2 is false and there exists some ensemble X such that $O < \tilde{O}$ for some value of $\left\langle D^2 \right\rangle > 0$. According to Lemma 1, O must eventually go to $\tilde{O}$ as we decrease $\left\langle D^2 \right\rangle$ towards zero; i.e. it must increase as $\left\langle D^2 \right\rangle$ decreases. This implies that there must be some value $\left\langle D^2 \right\rangle^*$ where $\left. \frac{\partial O}{\partial \left\langle D^2 \right\rangle} \right|_{\left\langle D^2 \right\rangle^*} < 0$, which contradicts Lemma 2. Therefore, $O \geq \tilde{O}$ is proved by contradiction■

Given Theorem 1 and Theorem 2, it is clear that $\log_2\left(\frac{n+1}{n}\right) \leq O \leq 1$.

Estimating $\left\langle D^2 \right\rangle$ from molecular simulations

To determine an estimate for $\left\langle D^2 \right\rangle$ at room temperature, we used data from representative trajectories within the Dynameomics project.[2-7] The Dynameomics database contains molecular dynamics simulations at 298K of at least 31ns for a relatively large selection of proteins corresponding to highly-populated structural folds. The all-atom simulations were conducted

with the *in lucem* molecular mechanics (*il*mm) program using the explicit solvent model F3C.[8-10] Ensembles were constructed for each protein from the trajectories by selecting 1000 structures, in 3ps intervals, from the first 30ns of simulation. We calculated the average pairwise MSD for each trajectory, $\left\langle D^2 \right\rangle_j$ (where j denotes the j^{th} trajectory), and estimated $\left\langle D^2 \right\rangle$ as the mean of these values after discarding any outliers; i.e., the 5 most and 5 least flexible proteins. This resulting value of $\left\langle D^2 \right\rangle = 2.75\,\text{Å}^2$ was used in all subsequent calculations. A comparison of the $\left\langle D^2 \right\rangle$ value calculated using only the first 22.5 ns of each simulation (2.6 Å^2) to its value calculated using the full 30 ns simulations (2.75 Å^2). The difference between the two values is less than 10%, thereby suggesting that the $\left\langle D^2 \right\rangle$ statistic had reasonably converged. To further demonstrate that these differences are not significant, we recalculated the order parameters for the Dynameomics data for a ±10% (0.275 Å^2) change in $\left\langle D^2 \right\rangle$, and found that changes of this magnitude had a minimal affect on the calculated order parameters as shown in Figure S1 below. In addition, we note that since the same value of $\left\langle D^2 \right\rangle$ is used to compute the order parameter value for each protein, the relative rank "ordering" of the proteins is not affected by the value of $\left\langle D^2 \right\rangle$.

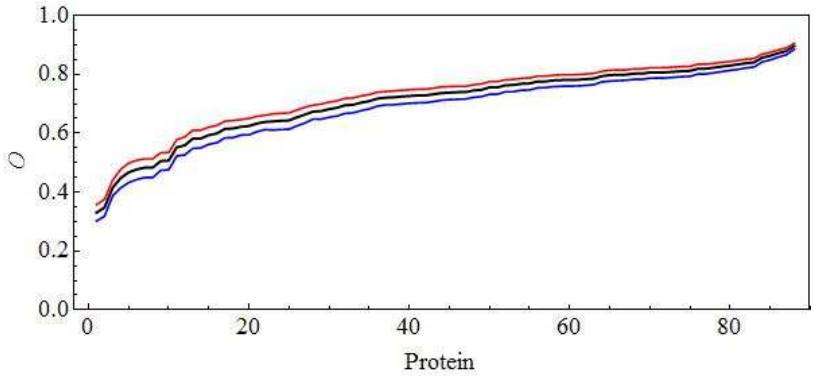

Figure S1. A 10% change in $\langle D^2 \rangle$ on calculated order parameters.

Derivation of an approximation for the order parameter from crystallographic B-factors

It is helpful to have an approximation to eq. (5) that can be calculated directly from X-ray crystallographic data. To accomplish this we first note that for any differentiable functions f, g of a random variable *X* one can perform a Taylor series expansion about $E[X]$ *to the lowest order* using the Law of Iterated Expectation to obtain:

$$E\left[f\left(E\left[g\left(X\right)\right]\right)\right] \approx f\left(g\left(E\left[X\right]\right)\right) \tag{6}$$

Next, we make use a recently derived relationship between the MSD and crystallographic B-factors:[11]

$$E\left[D^2\left(s_i, s_j\right)\right] \approx \frac{2}{N}\sum_{i=1}^{N}\frac{3B_i}{8\pi^2} \tag{7}$$

Combining eqs. (5)-(7) we obtain:

$$O \approx \log_2\left(1+\exp\left(\frac{-1}{2\langle D^2\rangle}\sum_{i=1}^{N}\frac{3B_i}{4\pi^2 N}\right)\right) \tag{8}$$

It is important to note that one limitation to eq. (8) is that it is based on the lowest order Taylor series approximation (eq. (5)(6)). To make sure that this approximation provides a reasonable level of accuracy, we compared order parameters calculated with eq. (5) to those calculated using the corresponding lowest order Taylor series approximation:

$$O \approx \log_2\left(1+\exp\left(\frac{-1}{2\langle D^2\rangle}\sum_{i=1}^{n} w_i \sum_{j=1}^{n} w_j D^2\left(s_i, s_j\right)\right)\right) \qquad (9)$$

for the ensembles obtained from the Dynameomics project.[2-7] As shown in Fig. S2, the lowest order approximation is generally accurate, particularly when $O > 0.5$ as was the case for all of the crystal structures discussed in the main text.

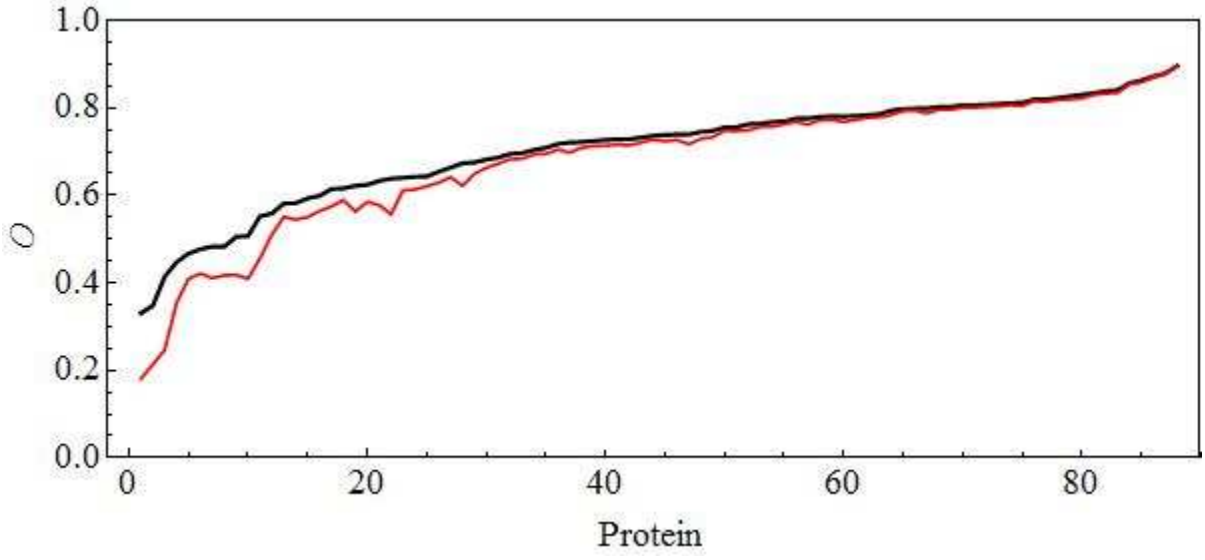

Figure S2. A comparison of the OP calculated using eq. (5) (black) and eq. (9) (red) for the Dynameomics ensembles.

List of PDB codes of proteins used from the Dynameomics project. 1wit, 3chy, 1ypi, 1ris, 1enh, 1shf, 1sac, 1ubq, 1mjc, 4icb, 1a6n, 1uu2, 1cun, 1ril, 1was, 1ep0, 2pth, 1qau, 1ebd, 1vid, 4wbc, 1dyn, 1jam, 2go0, 2giw, 1tmc, 1ifc, 1pma, 1snb, 1bp5, 1d1n, 1hh8, 1bs2, 1cok, 1ixa, 1vap, 1ier, 1jd1, 1dj1, 1ypr, 1bfd, 1eqk, 1gad, 1lbd, 1hgu, 1ab2, 1qaz, 1iad, 1b6b, 1fkb, 1ntn, 1nr2, 1byl, 1l8l, 2trc, 1fzw, 1bo9, 1ezg, 2hnp, 1r4v, 1bsg, 1p9g, 1d6t, 1esj, 1cvz, 1ceq, 1d0n, 11as, 1j1y, 1hcc, 1p99, 1cuk, 1u9a, 2lao, 1elp, 1bf0, 1uxc, 1php, 3grs, 1f8d, 1bhd, 1ddg, 1g5b, 2tgi, 1d8v, 1bqg, 1axj, 1agi, 1ihb, 1fzt.

References

(1)Kullback, S.; Leibler, R. A. ***Annals of Mathematical Statistics*** **1951,** ***22***, **79.**
(2)Beck, D. A.; Jonsson, A. L.; Schaeffer, R. D.; Scott, K. A.; Day, R.; Toofanny, R. D.; Alonso, D. O.; Daggett, V. ***Protein Eng Des Sel*** **2008,** ***21***, **353.**
(3)Benson, N. C.; Daggett, V. ***Protein Sci*** **2008,** ***17***, **2038.**
(4)Jonsson, A. L.; Scott, K. A.; Daggett, V. ***Biophys J*** **2009,** ***97***, **2958.**
(5)Kehl, C.; Simms, A. M.; Toofanny, R. D.; Daggett, V. ***Protein Eng Des Sel*** **2008,** ***21***, **379.**
(6)Simms, A. M.; Toofanny, R. D.; Kehl, C.; Benson, N. C.; Daggett, V. ***Protein Eng Des Sel*** **2008,** ***21***, **369.**
(7)van der Kamp, M. W.; Schaeffer, R. D.; Jonsson, A. L.; Scouras, A. D.; Simms, A. M.; Toofanny, R. D.; Benson, N. C.; Anderson, P. C.; Merkley, E. D.; Rysavy, S.; Bromley, D.; Beck, D. A.; Daggett, V. ***Structure*** **2010,** ***18***, **423.**
(8)Beck, D. A.; Daggett, V. ***Methods*** **2004,** ***34***, **112.**
(9)Levitt M, H. M., Sharon R, Laidig KE, Daggett V ***The Journal of Physical Chemistry B*** **1997,** ***101***.
(10)Levitt M, H. M., Sharon R, Daggett V. ***Computer Physics Communications*** **1995,** ***91***, **215.**
(11)Kuzmanic, A.; Zagrovic, B. ***Biophysical Journal*** **2010,** ***98***, **861.**