# Consequences of localized frustration for the folding mechanism of the IM7 protein - supplementary materials

**Ludovico Sutto** [*] [†],**Joachim Lätzer** [‡] [§],**Joseph A. Hegler** [‡] [§] ,**Diego U. Ferreiro** [‡] [§] , **and Peter G. Wolynes**[¶] [§]

[*]Department of Physics, University of Milano, via Celoria 16, 20133 Milano, Italy,[†] INFN, Sez. di Milano, Milano, Italy,[‡]Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, California 92093-0365, USA, and [§]NSF Center for Theoretical Biological Physics, San Diego, La Jolla, California 92093-0365, USA

Submitted to Proceedings of the National Academy of Sciences of the United States of America

## SI Text

**Frustration of the IM7–Colicin E7 Complex.** In nature, residues involved in binding have evolved to form a favorable binding interface. The local frustration of the IM7 chain in isolation and with the binding partner Colicin E7 appears in Fig. . Compared to the isolated IM7 monomer (Fig. *A*), region III in particular has fewer highly frustrated contacts ($F_{ij} < -1$, red) and more minimally frustrated ($F_{ij} > 0.78$, green) contacts when bound to Colicin E7 (Fig. *B*). Moreover, the IM7-Colicin E7 interface is dominated by minimally frustrated contacts. These observation support the idea that the frustrated regions become less frustrated upon binding to Colicin E7. Furthermore, the side chains of the three top residues selected for mutation (positions 49, 55, and 56) are located in the central region of the binding interface (Fig. *B*). Many contacts still exposed to solvent remain highly frustrated. Interestingly, the N-terminal region of the Colicin E7 (>400 aa in length) is absent from this crystal structure. A possible explanation for the remaining frustrated regions may be provided by interactions with this missing region of the Colicin E7 structure. Also, the biologically-relevant IM7-Colicin E7 oligomerization state may not be a dimer. In alternative oligomerization states, other regions of the IM7 surface may form additional interactions.

**Simulation Methods.** Two Hamiltonians with different contact energy terms are used to elucidate the folding of IM7. We first used an off-latice G$\bar{\text{o}}$-like model of the type introduced by Eastwood and coworkers (1) to investigate the role of topology on folding. We then performed molecular dynamics simulation with the AMW Hamiltonian (2). This Hamiltonian allows us to obtain details on the role of non-native contacts as well as water mediated interactions at a molecular level. The AMW Hamiltonian and the simulation protocols are described.

***Native topology-based simulations.*** We first studied the folding of IM7 with a native topology-based model yielding a perfectly funneled energy landscape.

$$\mathcal{H}_{G\bar{o}} = \mathcal{H}_{bb} + \mathcal{H}^{AM}_{G\bar{o}} \qquad [1]$$

Each term in this Hamiltonian does only depend on $C_\alpha$, $C_\beta$ and the backbone oxygen atoms, which are thus the only atoms to enter the dynamics. In this model the $\mathcal{H}_{bb}$ represents a generic backbone potential that assures correct backbone chemistry of the protein chain. The individual terms are explained in detail by Eastwood*et al.* (1) and are set to yield physically reasonable values. Secondary and tertiary structure are biased to be native-like with a potential whose minimum is at the native structure of IM7.

$$\mathcal{H}^{AM}_{G\bar{o}} = -\frac{\epsilon}{a_{G\bar{o}}} \sum_{i \leq j-3} \gamma_{G\bar{o}}[x(|i-j|)] \exp\left(-\frac{(r_{ij} - r^N_{ij})^2}{2\sigma^2_{ij}}\right) \qquad [2]$$

The sum over $ij$ runs over all unique pairs of carbon atoms used in the simulation separated in sequence by at least three residues. The interactions between $C_\alpha(i)$ and $C_\beta(j)$ is a Gaussian centered at the distance found in the native structure of IM7. The well widths is dependant on sequent separation, i.e. $\sigma_{ij} = |i-j|^{0.15}$. The strength of interaction is given by $\gamma_{G\bar{o}}[x(|i-j|)]$ and is scaled to yield equal energy in each proximity class (short range, medium range and long range interactions). The energy in each proximity class is normalized to yield a third of the native energy. Energetic heterogeneity is introduced by scaling the set of $\gamma$ values according to the set of weights determined by Goldstein*et al.* (3). Multiple trajectories with numerous unfolding/folding transitions were collected and analyzed using the histogram analysis method to calculate the free energy surface projected onto the fraction of native contacts ($Q_W$) and rmsd. The reaction coordinate $Q_W$ was previously demonstrated to accurately map to $P_{\text{fold}}$ at the resolution of $\Phi$-values for a funneled energy landscape (4). We incorporated nonadditivity, which implicitly accounts for sidechain and solvent interactions ordinarily absent from the pairwise additive model, into calculations of free energy profiles (5, 6).

***Molecular dynamics simulations with the AMW Hamiltonian.*** The AMW Hamiltonian is a coarse-grained, transferable potential designed to predict the global native fold of proteins from their sequence. The Hamiltonian is general and contains $20 \times 20$ contact potentials for direct and water mediated contacts that are modulated by the local environment. The basic mathematical form of the AMW is given by

$$\mathcal{H}_{AMW} = \mathcal{H}_{bb} + \mathcal{H}_{AM} + \mathcal{H}_{R_G} + \mathcal{H}_{contact} + \mathcal{H}_{water} + \mathcal{H}_{burial} \qquad [3]$$

and applies to a reduced set of coordinates of the heavy backbone atoms, $C^\alpha$, $C^\beta$ and oxygen. In this reduced description, the positions of the nitrogen and $C^{'}$ carbons are calculated assuming ideal protein backbone geometry. The Hamiltonian assures correct backbone chemistry and collapse of the protein. The functional forms of the individual terms of the Hamiltonian are explained in greater detail by Papoian*et al.* (2). We note that for residues <12 apart in sequence, the Associative Memory (AM) term applies while for residues separated by >12 in sequence, the contact potentials, $\mathcal{H}_{\text{contact}}$, $\mathcal{H}_{\text{water}}$ and $\mathcal{H}_{\text{burial}}$, apply. The AM term (1, 7) captures local structural folding propensities. When used in de novo structure prediction first one aligns the target sequence to memory proteins with the Local Hamiltonian (3), a sequence-structure alignment tool. The sequence is then threaded onto the memory proteins. These memory proteins

[¶]To whom correspondence should be addressed. Email: pwolynes@ucsd.edu

then determine the interactions for residues that are close in sequence therefore introducing a local secondary structure bias. In this study we use the IM7 crystal structure as the only memory protein. This assures that the local structure including secondary structure in the molecular dynamics simulations will be biased towards the local structure of the native state. The contact potential terms then predict the tertiary structure of the protein by flexibly assembling supersecondary structure elements.

To obtain the free energy landscape, constant temperature molecular dynamics runs were performed. Temperature is quoted in units of the native state energy of the AM and contact terms. The native state energy for a protein of $N$ residues is scaled in units of $\epsilon = \frac{E^{native}_{AM,contact}}{4N}$, which leads to define a reduced temperature as $k_BT = \epsilon\bar{T}$, where $\bar{T}$ is the temperature of the simulation. All other energy terms such as the backbone terms are scaled to yield physically reasonable interaction strengths. In a typical simulation run both a randomly unfolded structure as well as the x-ray structure were used as starting structures. Initial random velocities were assigned to the protein. For each temperature two sets of 20 trajectories were obtained (starting from either the native or random conformations). The length of each trajectory was $9\times10^6$ steps of approximate time length of 12-ns per step (8) resulting in 0.108-s long trajectories. In each of the runs 3,000 independent structural samples were obtained for analysis. For each sample the energies were recorded and the relevant order parameters were calculated. The free energy was then calculated at different temperatures using $F(Q_W, \mathrm{rmsd}) = -\bar{T}\cdot ln(P(Q_W, \mathrm{rmsd}))$ where $\bar{T}$ is the simulating temperature and $P(Q_W, \mathrm{rmsd})$ is the probability of finding a structure with given $Q_W$ and $\mathrm{rmsd}$ (9). These order parameters provide two different measures of the similarity to the crystal structure and both involves a sum over all (but nearest neighbour) pairs of $C^\beta$ or $C^\alpha$ atoms:

$$Rmsd = \sqrt{\frac{1}{\mathcal{N}}\sum_{i<j-1}(r_{ij} - r^{\mathrm{N}}_{ij})^2}, \qquad [4]$$

where $r^{\mathrm{N}}_{ij}$ is the $C^\beta$-$C^\beta$ distance between residues $i$ and $j$ in the native state,

$$Q_W = \frac{1}{\mathcal{N}}\sum_{i<j-1}\exp\left[-\frac{(r_{ij} - r^{\mathrm{N}}_{ij})^2}{\sigma^2_{ij}}\right], \qquad [5]$$

where $r^{\mathrm{N}}_{ij}$ is the $C^\alpha$-$C^\alpha$ distance between residues $i$ and $j$ in the native state, $\sigma_{ij} = |i-j|^{0.15}$ and the normalization $\mathcal{N} = (N-1)(N-2)/2$ is the number of non nearest neighbour pairs given the lenght N of the chain.

Thus a rmsd of $0\mathring{A}$ means the examined conformation is identical to the crystal structure whereas $Q_W$ ranges between 0 (completely unfolded) to 1 (native conformation) (1).

We define a non native contact as a $C^\beta$-$C^\beta$ pair whose distance in the crystal structure is $>9.5\mathring{A}$ and whose backbone distance is greater than four residues.

In order to calculate the folding time constant, we collected the distribution of times needed to make the transition from the unfolded state (rmsd $> 8.0\mathring{A}$) to the native state (rmsd $< 3.5\mathring{A}$). The exponential fit of this distribution leads to the folding time constant $\tau_f$.

The PDB ID for IM7 is 1AYI.

**Simulation results with funneled energy function.** The results for the simulations with the various funneled energy functions are described in detail in the main body of the article. The figure provided in the supplementary section shows results for simulations with a perfectly funneled Hamiltonian with various degrees of explicit cooperativity. When no cooperativity is present, IM7 folds as a downhill folder at all temperatures (see Fig. *A*). Stiffening the backbone chain results in two-state folding with a barrier of less than $1k_{\mathrm{B}}T$. When explicit cooperativity is added, IM7 clearly folds as a two-state folder (Fig. *B and C*). No intermediate is observed, even when contact heterogeneity is added.

## References


1. Eastwood MP, Hardin C, Luthey-Schulten Z, Wolynes PG (2001) *Ibm J Res Dev* **45**, 475–497.
2. Papoian GA, Ulander J, Eastwood M, Luthey-Schultes Z, Wolynes PG (2004) *Proc Natl Acad Sci USA* **101**, 3352–3357.
3. Goldstein RA, Luthey-Schulten ZA, Wolynes PG (1998) *Proc Natl Acad Sci USA* **95**, 4299–4302.
4. Cho SS, Levy Y, Wolynes PG (2006) *Proc Natl Acad Sci USA* **103**, 586–591.
5. Eastwood MP, Wolynes PG (2001) *J Chem Phys* **114**, 4702–4716.
6. Ejtehadi MR, Avall SP, Plotkin SS (2004) *Proc Natl Acad Sci USA* **101**, 15088–15093.
7. Hardin C, Eastwood MP, Prentiss M, Luthey-Schulten Z, Wolynes PG (2002) *J Comput Chem* **23**, 138–146.
8. Hardin C, Luthey-Schulten Z, Wolynes PG (1999) *Proteins* **34**, 281–294.
9. Sutto L, Tiana G, Broglia RA (2006) *Protein Sci* **15**, 1638–1652.