Supporting Information

Guseva et al. 10.1073/pnas.1620179114

SI Text

Simulations. To perform our stochastic simulations, we first needed to develop appropriate simulation code because of the large numbers of different molecular species that must be treated here. A description of the method, called the Expandable Partial Propensity Method (EPDM), and the corresponding C++ library can be found at https://github.com/abernatskiy/epdm (92). The challenge is to keep track of all of the molecular species and to search the full conformational spaces of each chain. This is nondeterministic-polynomial hard. We use the HP Sandbox algorithm (60, 97),* which is limited to maximum chain lengths of 25 monomers. To handle computational limitations, we restricted the total number of species to the level of a few thousands. We impose this limit by introducing a dilution parameter d: molecules are randomly removed from the system with probability $\propto d$. Physically, it represents molecules that diffuse out of the reaction volume. The total numbers of molecules within the reaction volume vary in the range of 10^2-10^4. We start our simulations with a small pool of monomers, usually fewer than 100 molecules. Here are the dynamical steps.

- Polymerization happens when monomers react with other monomers or polymers at a rate $\beta = 1$:

$$1\text{-mer} + n\text{-mer} \xrightarrow{\beta} (n+1)\text{-mer}. \qquad \textbf{[S1]}$$

- New monomers are imported into the system at high rate $a \gg 1$. From the point of view of the system, molecules appear out of nothing ($\emptyset$):

$$\emptyset \xrightarrow{a} H \text{ or } P. \qquad \textbf{[S2]}$$

- We assume that a fully unfolded chain can break at any internal site by hydrolysis. This happens with rate h per chain bond:

$$n\text{-mer} \xrightarrow{h} l\text{-mer} + (n-l)\text{-mer}. \qquad \textbf{[S3]}$$

- Typical values for the half-time for the hydrolysis of a bond under neutral conditions and room temperature are on the order of hundreds of years.†. Here, we explored a range of hydrolysis rates that are about $0.01-1$ of the polymerization rate. Hence, our model polymerization rates are on the order of days to years.
- We assume that the system becomes diluted at rate d. This has the practical purpose of limiting the total population of polymer in the system. We explored values of d from $\propto 0.01-1\beta$. Given the values of a that we used, it results in $\propto 10^2-10^4$ chains in the simulation volume

$$\text{anything} \xrightarrow{d} \emptyset. \qquad \textbf{[S4]}$$

- Folding and unfolding reactions happen much faster than the polymerization processes, with corresponding rate coefficients of $k_f \gg k_u \gg \beta$. The model considers folding to be a two-state process and does not take into account partially folded sequences because of the complexity of accounting for them:

$$\begin{gathered}\text{folded chain} \xrightarrow{k_u} \text{unfolded chain} \\ \text{unfolded chain} \xrightarrow{k_f} \text{folded chain}.\end{gathered} \qquad \textbf{[S5]}$$

- We used the most realistic values that we could obtain for these rates and for the folding free energies for proteins. We took E_{nat} from the HP model and known folding free energies from experimental data (100, 101), and we used the relationship (90)

$$\ln\left(\frac{k_f}{k_u}\right) = -\frac{\Delta G}{kT} = \frac{E_{nat}}{kT} - N \ln z, \qquad \textbf{[S6]}$$

- where z is the number of rotational dfs per peptide bond. To account for the difference between the 2D model and real 3D proteins, we calibrated the parameters taken from the literature to yield unfolding/folding rates that are meaningful in the context of the other rates in our model: folding is much faster than growth and for any of the sequence in our pool $k_f/k_u \in (10^2, 10^4)$ (100, 101) for 3D proteins. Because the literature models are only mean field, averaged over sequences, and to retain sequence dependence here, we set the unfolding rate of all sequences to the average for their lengths and assigned all of the sequence dependence to k_f. Therefore, we used

$$\begin{aligned}k_u &= \exp[12 - 0.1\sqrt{N} - E_H(0.5N + 1.34)], \\ k_f &= k_u \exp(\Delta G).\end{aligned} \qquad \textbf{[S7]}$$

- The model is not sensitive to varying these parameters over a wide range. We use $E_h \approx (1-2)kT$, and therefore, $k_{unf} \approx 10^2$, which leads to a range of unfolding rates from one unfolding per hour to one unfolding per day. Folding rates vary from a reaction per hour to a reaction per fraction of a second.
- The catalytic step is (only H monomers can be added to the sequences catalytically)

$$\text{Catalyst} + H + \underbrace{\cdots HH}_{l-1} \xrightarrow{\beta_{cat}} \text{Catalyst} + \underbrace{\cdots HHH}. \qquad \textbf{[S8]}$$

- The rate enhancement is $\beta_{cat} = \beta \cdot \exp(E_h \cdot n_c/kT)$, where hydrophobic sticking energy is e_H, and the number of contacting hydrophobes is n_c, which varies in the range from three to six. With the hydrophobic energies of $e_H = 1 - 2kT$, this gives catalysis rates around hours to days per reaction. Because the EPDM supports only binary reactions, we divided the reaction above into two steps: interaction of catalyst with a monomer with rate β and the reaction of this complex with a polymer that has the rate β_{cat}.

For each trajectory, we collected statistics only after the system reached an unchanging steady state. To explore the stochasticity, we repeated every simulation for 30 times for every experiment. We ran all of the simulations for $140s$ of internal simulation time, during which $10^6 - 10^9$ individual reactions had occurred. We took measurements every $10^{-6}s$. For all of the trajectories, steady-state behavior was reached no longer than $40s$ after the start of a simulation. Thus, we considered only the last $100s$ (1 million recordings) for each simulation. All of the data points that we used in the figures are averages over these recordings.

For all of the experiments below, we used the following parameters:

i) $\beta = 1$.
ii) $E_h = 2kT$.

*A Python implementation and description can be found at hp-lattice.readthedocs.org/en/latest/.

†The hydrolysis rate constants of oligopeptides in neutral conditions are of the order of $10^{-11} - 10^{-10}$: $1.310^{-10} M^{-1}s^{-1}$ for benzoylglycylphenylalanine ($t_{1/2} = 128\,y$) (98), $6.310^{-11} M^{-1}s^{-1}$ ($t_{1/2} = 350\,y$) for glycylglycine, and $9.310^{-11} M^{-1}s^{-1}$ for glycylvaline (99).

iii) $z = 1.2$.
iv) $a = 1000$.

Values of $a \ll 1000$ or $a \gg 1000$ are problematic, having numbers of sequences or populations either too high to calculate or too low to draw conclusions.

v) $h = d = 0.1$.

When $3d \lessapprox h \leq \beta$, hydrolysis dominates, and without catalysis, there is an explosion of short sequences.

When $3h \lessapprox d \leq \beta$, hydrolysis is unphysically small, and therefore, nothing limits the growth of longer sequences, even in the absence of catalysis.

When $0.05 \lessapprox d \approx h \lessapprox 0.5$, the forces of dilution and hydrolysis are relatively balanced, and populations are neither too small nor too large.

In Silico Experiments. The simulations were performed on the Laufer Center's computing cluster of central processing units. We performed the following computational experiments:

Experiment 1: Does our bare polymerization reproduce the Flory distribution? We started simulations with a small pool of chains up to 3-mers. To calculate the length distributions, we calculated for each trajectory the average population of every sequence over time over all recordings after 40 s, resulting in 1 million time steps. Then, we summed all of the populations of a given length, obtained total populations for all n-mers $n \in [1, 25]$, and then, computed every population as

$$p_n = \frac{\sum \text{all } n\text{-mers}}{\sum \text{total population}}, \quad \textbf{[S9]}$$

giving probability of finding an n-mer of a randomly chosen molecule in the system.

Experiment 2: What is the effect on the distribution of just HP folding? We started with the same initial population as in experiment 1. However, now, we introduce the hydrophobic energy $e_h = 2kT$. To calculate the result in length distribution, we computed the average population of every sequence for each trajectory over time over all of the recordings after 40 s, resulting in 1 million time steps.

Experiment 3: What is the effect on the distribution of both folding and catalysis? In addition to folding in this in silico experiment, we also accounted for the pairwise contact interactions between two proteins, with the parameters as indicated above. We explored ranges of parameters. We observed significant stability of the length distribution toward change of h and d in the range $0.05 \lessapprox d \approx h \lessapprox 0.5$. The distributions that we observe are quite sensitive to the choice of hydrophobic energy, as expected for chemical reactions, since this enters into the exponent of the rate expression. In the generally physical range of $e_h = 1 - 3kT$, we observe a bending of the Flory distribution, as noted in the text.