Self-Assembly Kinetics of Nanoscale Fused β-Solenoid Protein Units

Talia Sopp

University of Puget Sound

Rachel Baarda and Daniel Cox

Department of Physics, University of California, Davis

(Dated: May 21, 2017)

Abstract

β-solenoid protein (BSP) backbones twist helically to form coils of β-sheets. BSPs are mechanically robust, are easily customizable, and can self-assemble into complexes at room temperature. BSPs can be fused to small symmetric oligomers to create protein lattices with potential industrial applications ranging from synthetic antibodies to scaffolding nanomaterials. To assess the feasibility of creating such lattices, we modeled the formation of one of the simplest cases (a single hexagon) in *E. coli*. The hexagon is composed of trimer subunits in which two of the monomers have BSPs fused to them; six of these subunits form a hexagon. We modeled the formation of these subunits in *E. coli* as a series of diffusion-controlled reactions. We used two models to estimate the amount of this product and others over time: the deterministic reaction-rate theory and the stochastic Gillespie Method. Both showed that we could expect about 300 hexagon subunits to form in 25 minutes in one cell. We conclude that creating our hexagon BSP structure in *E. coli* is feasible. Our results will inform the experimental production of the hexagonal BSP structure. Additionally, we can apply the simulation method we developed to more complex protein lattices.

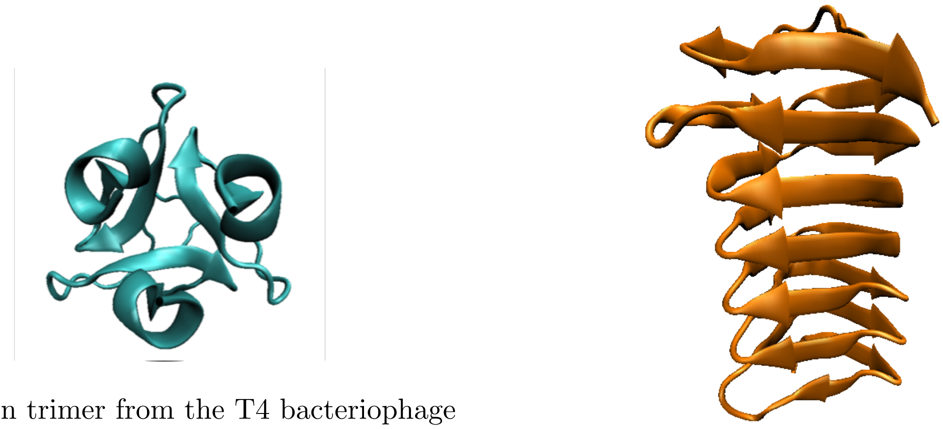
INTRODUCTION

One of the key goals of nanotechnology is to make things that are customizable on the atomic scale. Biomaterials have promise in nanotechnological applications because they have low toxicity and self-assemble at room temperature. DNA has been successfully used to build scaffolds that organize gold nanoparticles in two dimensions [1], to create 2D and 3D shapes ranging from a smiley face to a box that can be locked and unlocked for drug delivery, and to create nanomechanical devices [2]. However, there are several problems with DNA nanotechnology that hinder it from being used industrially. Nearly all DNA nanostructures are created by chemical synthesis (rather than in vivo processes), which is expensive and provides low yield of the desired structures [3]. At the moment, real applications of DNA nanostructures are quite rare.

Proteins have been explored as an alternative biomaterial for nanostructures. While synthesizing engineered proteins also requires synthetic DNA, that DNA is inserted in bacterial cells in culture, and from just one copy of a synthetic gene, many copies of the protein are produced. Synthetic protein production is therefore more scalable. While DNA is composed of 4 structurally similarly nucleic acids, proteins are built of 20 amino acids with widely varying properties. This gives proteins more conformational variability and functional versatility than nucleic acids [4].

β-solenoid proteins (BSPs) are composed of coiled beta sheets. Beta sheets are strands of protein held together by hydrogen bonds between the protein backbone. BSPs have functions in nature ranging from antifreeze proteins to prions. Their long, uniform sides can serve a variety of purposes. In the spruce budworm antifreeze protein (SPAFP), one face has a 2D array of threonine that hydrogen bonds to the surfaces of ice crystals, preventing further growth [5]. Inspired by this, our lab has investigated possibilities for modifying sides for other purposes, ranging from nanoparticle templating to creating synthetic antibodies.

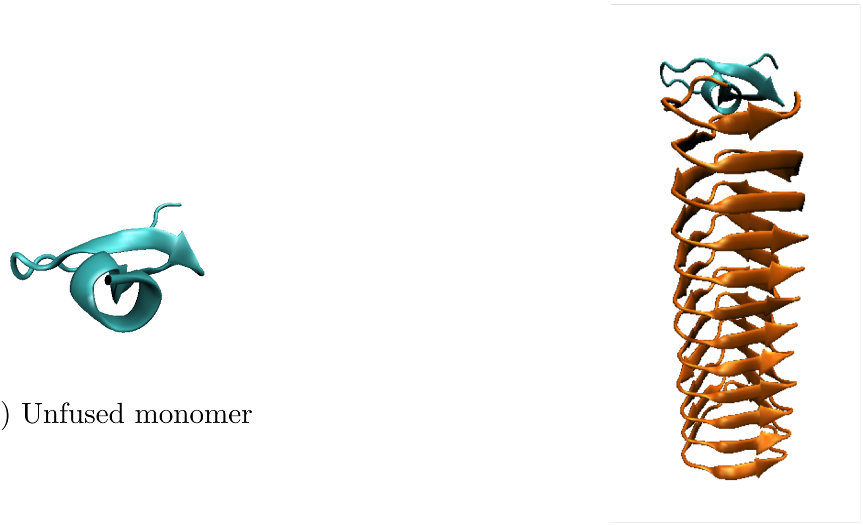
Because BSPs are mechanically robust [6], [7] and highly customizable, we propose using them as building blocks for nanoscaffolds. By fusing BSPs to small, symmetric oligomers, we can create a variety of scaffolds. One scaffold of interest is a 2D lattice of repeating hexagons. To investigate whether such a scaffold could be produced with significant yield, we computationally modeled the formation of a simpler but related form (a hexagon) composed of BSPs and symmetric trimers in *E. coli.* Since creating this hexagonal protein experimentally is



(a) Foldon trimer from the T4 bacteriophage

(b) Spruce budworm antifreeze protein (SBAFP)

FIG. 1: Proteins used to create monomer building blocks.



(a) Unfused monomer

(b) Fused monomer

FIG. 2: Monomer building blocks. Note that the SBAFP has been modified here to be longer than in its natural state.

very expensive and time intensive, the purpose of this computational project was to assess the feasibility of creating the hexagon protein and then to determine how experimental production might be maximized.

BACKGROUND

Structurally, this project required two types of proteins: a three-part connector and a long rod. For the three-part connector, we used the trimeric foldon domain of the protein fibritin from the T4 bacteriophage (“foldon” for short), shown in figure 1a (1RFO and 4NCU in PDB; see Appendix A for a distinction between the two PDB codes). For the long rod, we used the spruce budworm antifreeze protein, shown in figure 1b (SBAFP, 1M8N in PDB), which we then modified to add length. Though there are many proteins that we could have used to fulfill our structural requirements, we chose these specific proteins because they have been well characterized in the literature, are easy to produce in *E. coli* at room temperature, and have been successfully produced by our collaborators in the Toney biochemistry lab. From these two proteins, we create two monomers that will serve as the building blocks for our hexagons. One of these is a monomer from the foldon (figure 2a). The other is the foldon monomer with the SBAFP covalently bonded to its C-terminus (figure 2b). We will henceforth refer to the lone foldon monomer as the “unfused monomer,” and the foldon monomer with the SBAFP as the “fused monomer.”

Creating the monomers in the cell happens in the following manner (note that this was not actually done in this computational project). First, genes coding for the fused and unfused monomers are created. Then, those genes are inserted into *E. coli*. The *E. coli* cell transcribes the genes to RNA, then translates the RNA to create the fused and unfused monomer proteins. In the cytosol, the monomers non-covalently bind to each other to form various dimer and trimer products (figure 3). The hexagon (figure 4) is composed of trimer subunits, each with two fused and one unfused monomer (referred to as the uff trimer). When looking at figure 3, note that the uf and fu dimers are distinct species. This is because the foldon monomer has rotational symmetry but not reflectional symmetry.

Creating the hexagons requires two different *E. coli* cultures, as shown in figure 5. In one, the fused monomer is synthesized with the N-terminus of the SBAFP facing outwards (that is, not bonded to the foldon). In the other, it is the C-terminus. The uff trimers are isolated from each of these cultures, then combined. The N-termini and C-termini form peptide bonds. Having two separate cultures would allow us to maximize our yield of hexagons for the following reason. If the N-terminus and C-terminus monomers were created in the same culture, then all of the different protein products (shown in figure 3) could peptide bond to

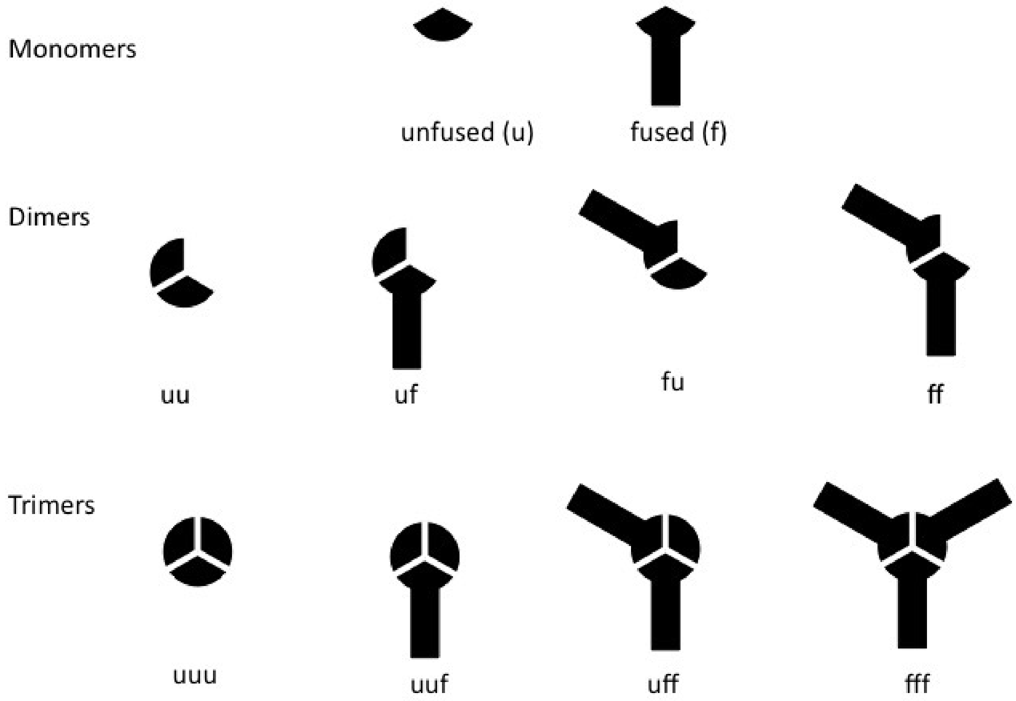


FIG. 3: Ten possible products. The uff trimer is the subunit that makes up the hexagon.

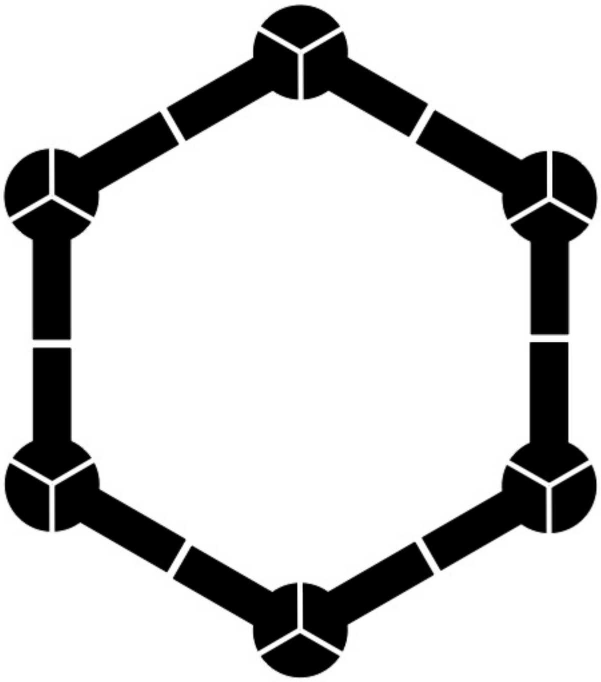


FIG. 4: The BSP hexagon, composed of uff subuints

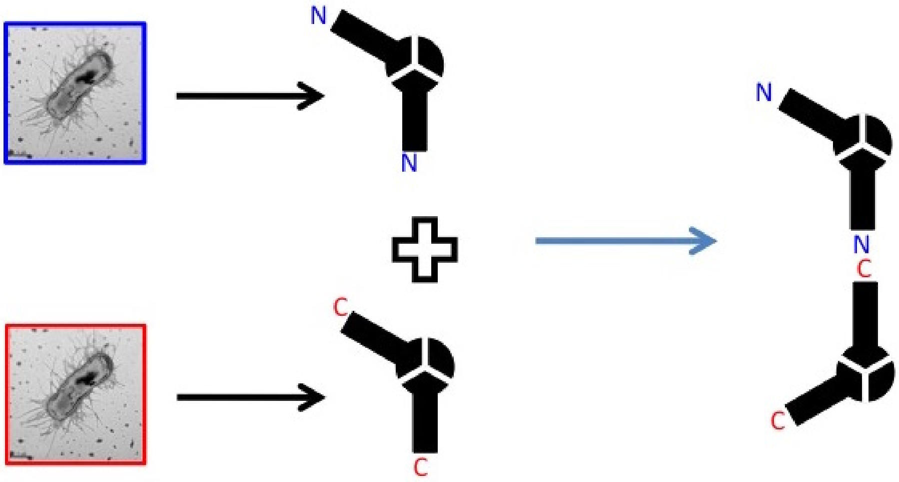


FIG. 5: Two different cultures of *E. coli* are required to make the hexagons. In one culture (blue), the fused monomers have the N-terminus facing outwards, while the other (red) has the C-terminus facing outwards. The uff monomers are isolated from each culture and then combined; they peptide bond to create the hexagons.

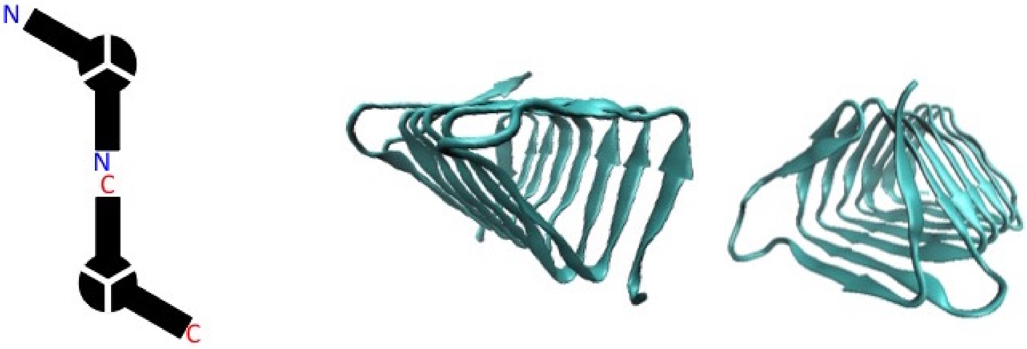


FIG. 6: On the left, a conceivable but unrealizable peptide binding orientation of two uff trimers is shown. The reason for why such an orientation is possible is shown on the right; the binding ends of the BSP do not align and a continuous BSP cannot form.

each other. With all the different peptide bonding options, it is unlikely that six uff subunits would manage to bind to each other rather than to other proteins.

One can imagine a variety of ways that the uff subunits could bind to form shapes other than the planar hexagon–squiggly lines and loops that are not restricted to the plane of the page. However, the shape of the binding site at the free end of the BSPs allows only the hexagon to form. As you can see in figure 6, two BSPs that are not aligned in the proper orientation will be unable to bind to form a continuous BSP. Though one can imagine three ways one could orient the triangular prism BSPs shown in figure 6 to make them align to make one continuous triangular prism, the locations of the binding sites themselves are such that only one of these configurations is allowed–the configuration that creates the hexagons. Further specifications about this matter are beyond the scope of this paper.

To guide hexagon production, we must determine how much uff we can expect the *E. coli* to produce over what time range and maximize the uff production rate.

MAXIMIZING UFF PRODUCTION

The cell can be manipulated to produce the fused monomers and unfused monomers at different rates. We would like the adjust these rates to maximize uff production. If p_f is the proportion of monomers that are fused and p_u is the proportion of monomers that are unfused (and $p_f + p_u = 1$), then the proportion of trimers that are of the uff configuration is:

$$P_{uff} = 3p_f^2(1 - p_f)$$

(Note: the expected term, $p_f^2(1 - p_f)$, is multiplied by 3 because there are three different combinations that form the uff trimer: the u monomer and ff dimer, f and uf, and f and fu).

$$\frac{\mathrm{d}P_{uff}}{\mathrm{d}p_f} = 0 \text{ when } p_f = 2/3 \text{ and } p_u = 1/3$$

Therefore, $p_f : p_u$ for maximal uff production is 2:1. At this ratio, four of every nine trimers will be uff trimers.

Now that we have the fused to unfused production *ratio*, we must now estimate the actual production *rate* for each monomer. Since the factor that can be controlled in experiment is the number of genes inserted into *E. coli*, we need to know what f:u gene ratio results in a 2:1 production rate ratio, and use this to estimate the production rate of each protein.

In their modified forms in the fused monomer, SBAFP and the foldon monomer are 180 and 28 amino acids long, respectively. Assuming that there are no non-coding regions in the genes, we multiply the number of amino acids by three to get the following gene lengths:

Species	protein length (AAs)	gene length (BPs)
SBAFP	180	540
Foldon monomer	28	84
u	28	84
f	208	624

(Note: these are the lengths for the foldon monomer and SBAFP molecules as they are modified for use in our products, not the lengths of the naturally-occurring proteins as given in their PDB files).

The maximum rate of transcription in *E. coli* is about 40-80 nucleotides per second, and the maximum rate of translation is roughly 60 nucleotides per second [8]. Since these rates are approximately equal, we can define either as the rate-limiting step [8]. Taking the lower end of the range to be conservative, we used 40 nucleotides per second as the rate of transcription, which we defined as our rate-limiting step. If we assume that our genes of interest are continually transcribed, we arrive at the following rates of transcription for our genes.

Species	Rate of transcription (gene/second)
u	0.476
f	0.0641

Let us turn our gene transcription rates into protein production rates. If there is only one copy of the gene in the genome, then only one RNA polymerase (and, subsequently, one ribosome) can be working on it at a time. Then, each time the gene is transcribed, one molecule of the protein is produced. We can therefore change the units of our rate of transcription from genes per second to molecules per second. Then, the unfused monomer production rate for the entire cell is:

$$(0.476 \text{ molecules/s}) \times (1 \text{ mole}/6.022 * 10^{23} \text{ molecules}) \times (1/10^{-15}\text{L}) \approx 7.9 \times 10^{-4}\mu\text{M/s}$$

where 10^{-15} L, or $1\mu\text{m}^3$ is the approximate volume of an *E. coli* cell [8].

For the fused monomer, we get a production rate of approximately $1.1 \times 10^{-4}\mu$M/s for one gene. However, we want production of the fused monomers to be twice that of the unfused monomers. We must multiply the fused monomer production rate by 15 to achieve our 2:1 production ratio, signifying that we require about 15 times as many copies of the fused gene as the unfused. This makes intuitive sense: a larger gene (the fused gene) takes a longer time to transcribe and translate, so you need many more copies of the gene to achieve a 2:1 production ratio. With the 15:1 gene ratio, our final production rates are:

Species	production rate (μM/s)
u	7.9×10^{-4}
f	16×10^{-4}

ESTIMATING ACCUMULATION OF PRODUCTS OVER TIME

In order to determine how much of the uff trimer we can expect the cell to produce, we considered all possible reactions involving the unfused and fused monomers. These are as follows:

$$\begin{gathered}
\star \longrightarrow \mathrm{u} \\
\star \longrightarrow \mathrm{f} \\
\mathrm{u} + \mathrm{u} \longrightarrow \mathrm{uu} \\
\mathrm{u} + \mathrm{f} \longrightarrow \mathrm{uf} \\
\mathrm{u} + \mathrm{f} \longrightarrow \mathrm{fu} \\
\mathrm{f} + \mathrm{f} \longrightarrow \mathrm{ff} \\
\mathrm{uu} + \mathrm{u} \longrightarrow \mathrm{uuu} \\
\mathrm{uu} + \mathrm{f} \longrightarrow \mathrm{uuf} \\
\mathrm{uf} + \mathrm{u} \longrightarrow \mathrm{uuf} \\
\mathrm{fu} + \mathrm{u} \longrightarrow \mathrm{uuf} \\
\mathrm{uf} + \mathrm{f} \longrightarrow \mathrm{uff} \\
\mathrm{fu} + \mathrm{f} \longrightarrow \mathrm{uff} \\
\mathrm{ff} + \mathrm{u} \longrightarrow \mathrm{uff} \\
\mathrm{ff} + \mathrm{f} \longrightarrow \mathrm{fff}
\end{gathered}$$

Note that the first two reactions are the production of the monomers from DNA. Each of these reactions moves forward at a different rate. We do not include the reverse reactions (dissociations) because the rates of these are so small that they are negligible; we can assume that the association reactions are irreversible. We used methods that considered all of these association reactions to estimate how the concentration of each of the 10 species varies over time. The first method used the deterministic reaction rate equations and the second used the stochastic Gillespie method.

Method 1: Deterministic reaction rate equations

Let us consider an example of a reaction rate equation, say, the reaction between the unfused monomer and the fused monomer to produce the uf dimer. The equation for this reaction is an ordinary differential equation that relates the time derivative of the concentration of the product to the concentration of each of the reactants and the rate constant, k.

$$\frac{\mathrm{d}[uf]}{\mathrm{d}t} = k_{uf}[u][f] \tag{1}$$

An increase in the product corresponds to a proportional decrease in the reactants, so

$$\frac{\mathrm{d}[f]}{\mathrm{d}t} = \frac{\mathrm{d}[u]}{\mathrm{d}t} = -\frac{\mathrm{d}[uf]}{\mathrm{d}t} = -k_{uf}[u][f] \tag{2}$$

Equation 1 is not a complete equation for the rate of change of concentration of the uf species. Looking at our reactions listed in the previous section, we see that there are two other reactions involving the uf species. The corresponding reaction rate equations for these are,

$$\frac{\mathrm{d}[uuf]}{\mathrm{d}t} = k_{ufu}[uf][u] \tag{3}$$

which corresponds to

$$\frac{\mathrm{d}[uf]}{\mathrm{d}t} = -k_{ufu}[uf][u] \tag{4}$$

and

$$\frac{\mathrm{d}[uff]}{\mathrm{d}t} = k_{fuf}[uf][f] \tag{5}$$

corresponding to

$$\frac{\mathrm{d}[uf]}{\mathrm{d}t} = -k_{fuf}[uf][f] \tag{6}$$

By adding all of our terms for the rate of change of concentration of uf, we arrive at a final equation:

$$\frac{\mathrm{d}[uf]}{\mathrm{d}t} = k_{uf}[u][f] - k_{ufu}[u][uf] - k_{fuf}[f][uf] \tag{7}$$

Repeating this process for each species involved, we end up with a set of 10 coupled nonlinear first order ODEs:

$$\frac{\mathrm{d}[u]}{\mathrm{d}t} = c_u - 2k_{uu}[u]^2 - 2k_{uf}[u][f] - k_{uuu}[u][uu] - k_{ufu}[u]([uf]+[fu]) - k_{uff}[u][ff] \tag{8}$$

$$\frac{\mathrm{d}[f]}{\mathrm{d}t} = c_f - 2k_{ff}[f]^2 - 2k_{uf}[u][f] - k_{uuf}[f][uu] - k_{fuf}[f]([uf]+[fu]) - k_{fff}[f][ff] \tag{9}$$

$$\frac{\mathrm{d}[uu]}{\mathrm{d}t} = k_{uu}[u]^2 - k_{uuf}[f][uu] - k_{uuu}[u][uu] \tag{10}$$

$$\frac{\mathrm{d}[uf]}{\mathrm{d}t} = k_{uf}[u][f] - k_{ufu}[u][uf] - k_{fuf}[f][uf] \tag{11}$$

$$\frac{\mathrm{d}[fu]}{\mathrm{d}t} = k_{uf}[u][f] - k_{ufu}[u][fu] - k_{fuf}[f][fu] \tag{12}$$

$$\frac{\mathrm{d}[ff]}{\mathrm{d}t} = k_{ff}[f]^2 - k_{uff}[u][ff] - k_{fff}[f][ff] \tag{13}$$

$$\frac{\mathrm{d}[uuu]}{\mathrm{d}t} = k_{uuu}[u][uu] \tag{14}$$

$$\frac{\mathrm{d}[uuf]}{\mathrm{d}t} = k_{ufu}[u]([uf]+[fu]) + k_{uuf}[f][uu] \tag{15}$$

$$\frac{\mathrm{d}[uff]}{\mathrm{d}t} = k_{fuf}[f]([uf]+[fu]) + k_{uff}[u][ff] \tag{16}$$

$$\frac{\mathrm{d}[fff]}{\mathrm{d}t} = k_{fff}[f][ff] \tag{17}$$

$$ \tag{18}$$

Note here that c_f and c_u are production rates–that is, the rates at which the concentrations of the monomers increase due to their production from DNA. Recall from the previous section that our production rates are:

Species	production rate, c (μM/s)
u	7.9×10^{-4}
f	16×10^{-4}

Note the distinction between seemingly similar rate constants. Since the combination of u and f to produce uf and fu involved the same reactants (and the two binding sites on each monomer are favored equally), the rate constant k_{uf} is the same for both reactions. However, uff can be created by the combination of f and uf or fu, or u and ff. The uf/fu reactions with f have one rate constant, k_{fuf}, while the reaction between u and ff has another rate constant, k_{uff}.

Rate constants

To make use of these equations, we had to determine the values of each of the rate constants. Ideally, one would determine the rate constant for each reaction experimentally. However, since most of these complexes have yet to be produced experimentally, we used models to estimate the rate constants. From the paper "Very Fast Folding and Association of a Trimerization Domain from Bacteriophage T4 Fibritin" by Güthe et al, the experimental bimolecular rate constants for foldon dimer and trimer formation (that is, uu and uuu production) are $1.9(\pm 0.5) \times 10^6\ \mathrm{M}^{-1}\mathrm{s}^{-1}$ and $5.4(\pm 0.3) \times 10^6\ \mathrm{M}^{-1}\mathrm{s}^{-1}$, respectively, for association experiments performed in water [9]. For the other reactions, we attempted several different methods for approximating rate constants.

Our association reactions are diffusion-limited. This means that if the two binding sites come into contact, they will always bind. The rate-limiting factor is the time it takes for two proteins to diffuse through the cell to come into proximity of each other, and to rotate so that their binding sites align. We modeled the reaction between two of our protein species as a reaction between two spheres. Though the spherical representation obviously does not reflect the physical reality of the situation, the more precise models we considered using were deemed too complicated and time-intensive relative to the prospective gains for this project. To judge the success of a model, we compared the values it gave for k_{uu} and k_{uuu} in water to the experimentally known constants.

We modeled the association of two proteins as a reaction between two spheres, each with a specific binding patch (see figure 7). The equation we used for estimating the diffusion-controlled reaction rate constant for the association of molecule A and molecule B is:

$$k_{AB} = \pi(D_A + D_B)(r_A + r_B)\sin\theta_A \sin\theta_B \sin\left(\frac{\theta_A + \theta_B}{2}\right) \tag{19}$$

where D is the Stokes-Einstein diffusion coefficient

$$D = \frac{k_B T}{6\pi\eta r} \tag{20}$$

Equation 19 here is listed as equation 6 in the cited paper by Berg and von Hippel [10]. Note that the equation in that paper uses the small angle approximation; we did not use it here because our largest angle was $\frac{\pi}{6}$.

The following values were required: viscosity of the fluid (η), and diffusion coefficient (D),

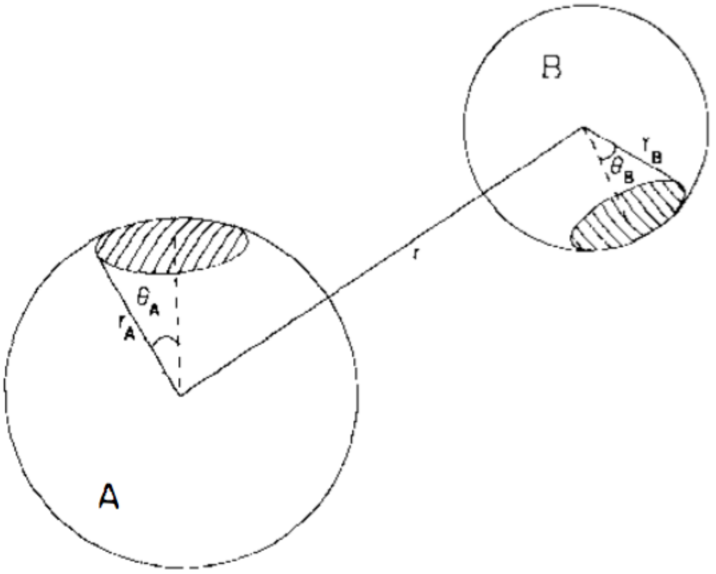


FIG. 7: Binding patch geometry for two molecules with different binding patch sizes. [10]

radius (r), and binding patch angle (θ) for each molecule, assuming spherical molecules. We needed the viscosity of water in order to compare our calculations to the experimental rate constants found in the Güthe et al. paper [9], which were done in water. We needed the viscosity of the cytoplasm of *E. coli* to model the intracellular environment for our reactions.
$\eta_{water} \approx 1.0$ mPa·s
$\eta_{cytoplasm} \approx 25$ mPa·s [11]
Using the radius of gyration found in VMD (Visual Molecular Dynamics, a molecular modeling and visualization program) as an estimate of the molecular radius, we get the following radii for monomers and dimers:

Species	radius (m)
u	0.939×10^{-9}
f	2.25×10^{-9}
uu	1.06×10^{-9}
uf/fu	2.57×10^{-9}
ff	3.46×10^{-9}

Next, we found the diffusion coefficient by using the Stokes-Einstein relation (see equation 20), where k_B is Boltzmann's constant and $T = 293$ K.
For u and uu in water:

Species	D (m^2/s)
u	2.29×10^{-10}
uu	2.03×10^{-10}

For the monomers and dimers in the cytoplasm of *E. coli*:

Species	D (m^2/s)
u	9.14×10^{-12}
f	3.81×10^{-12}
uu	8.11×10^{-12}
uf/fu	3.34×10^{-12}
ff	2.48×10^{-12}

Binding patch size

Since the monomers are held together by hydrogen bonds, we used the cross sectional area of a hydrogen bond as the radius of the binding patch size for two monomers binding. We found inspiration from the Torshin et al paper [12] for estimating the binding patch size for a hydrogen bond. Since the angle between the donor atom, bound hydrogen, and accepting atom must be between 90 and 180 degrees, and the accepting atom could be located at any rotational location around the axis formed by the donor atom and the hydrogen, we ended up with a hemispherical shell with the radius of a hydrogen bond centered at the hydrogen atom, anywhere on which the accepting atom could bind to form the hydrogen bond. This hemispherical shell is our binding patch area. However, since our model requires us to approximate our binding patch as a 2D circle (see again figure 7), we reduced our binding patch area to πr^2, where r is the length of a hydrogen bond. We found the average length for the hydrogen bond between monomers by using VMD to measure the length of hydrogen bonds between monomers in 10 different frames (that is, at 10 different moments while the molecule is jiggling). Therefore, for monomers,

$$r_{patch,mono} = 0.185 \text{ nm}$$

For dimers, the reaction patch is larger, since two hydrogen bonds form instead of one. Taking a cue from Xie et al, 2016 [13], we said that for the uu dimer, the reaction patch is about 1/6 of a sphere, leading to $\theta = \frac{\pi}{6}$. Using $r_{patch} = r_{uu} \sin\theta$, this corresponds to a reaction patch size of

$$r_{patch,dim} = 0.530 \text{ nm}$$

Because the binding patch size is the same for all of the dimers, the same binding patch radius was used for all dimers. Using $\theta = \sin^{-1}(r_{patch}/r_x)$ (where x is the relevant species), we arrived at the following binding patch angles:

Species	θ (radians)
u	0.198
f	0.0821
uu	0.524
uf/fu	0.207
ff	0.154

Plugging the appropriate values into equation 19 for the formation of the uu dimer and uuu trimer in water, we arrive at the following rate constants, compared to the experimental values:

Species	Calculated k ($\text{M}^{-1}\text{s}^{-1}$)	Experimental k ($\text{M}^{-1}\text{s}^{-1}$)
uu	1.23×10^7	1.90×10^6
uuu	5.66×10^7	5.40×10^6

The calculated rate constants are an average of 12.5 times larger than the experimental rate constants. Since the only difference between our rate constants in water and in *E. coli* is viscosity (and we are using experimental values for viscosity), we expect our calculations in *E. coli* to be 12.5 times too large as well. It is difficult to discern precisely why our calculations are larger than the experimental values because there are many layers of approximation involved here, from modeling the molecules as spheres to estimating a circular binding patch size. To better approximate our rate constants, we divided the calculated values for our rate constants by 12.5 to arrive at our final rate constants.

Species	Calculated k ($\text{M}^{-1}\text{s}^{-1}$)	Experimental k ($\text{M}^{-1}\text{s}^{-1}$)
uu	1.54×10^6	1.90×10^6
uuu	7.07×10^6	5.40×10^6

To translate these to reactions in *E. coli*, we simply substitute the viscosity of the cytoplasm of *E. coli* for the viscosity of water to arrive at the following rate constants:

Reaction	k symbol	k value ($\mathrm{M}^{-1}\mathrm{s}^{-1}$)
u + u	k_{uu}	6.2×10^4
u + f	k_{uf}	2.2×10^4
f + f	k_{ff}	4.5×10^3
u + uu	k_{uuu}	2.8×10^5
f + uu	k_{uuf}	1.1×10^5
u + uf and u + fu	k_{ufu}	8.4×10^4
f + uf and f + fu	k_{fuf}	2.0×10^4
u + ff	k_{uff}	6.4×10^4
f + ff	k_{fff}	1.3×10^4

By inserting our calculated rate constants into our set of 10 ODEs and computationally integrating them over time, we graphed how the concentration of each species varies over time, shown in figure 8. The monomers are constantly being produced from DNA and consumed by the dimer- and trimer-forming reactions; the dimers are being formed from monomers and consumed to form trimers; the trimers are being formed from monomers and dimers and are not being consumed in any reaction. Therefore, the concentration of monomers and dimers should reach equilibrium concentrations where the producing and consuming reactions balance out and the trimers should continue to increase indefinitely. This trend is seen in figure 8. Additionally, uff ultimately becomes the majority species, as we expected from optimizing our production rate ratios.

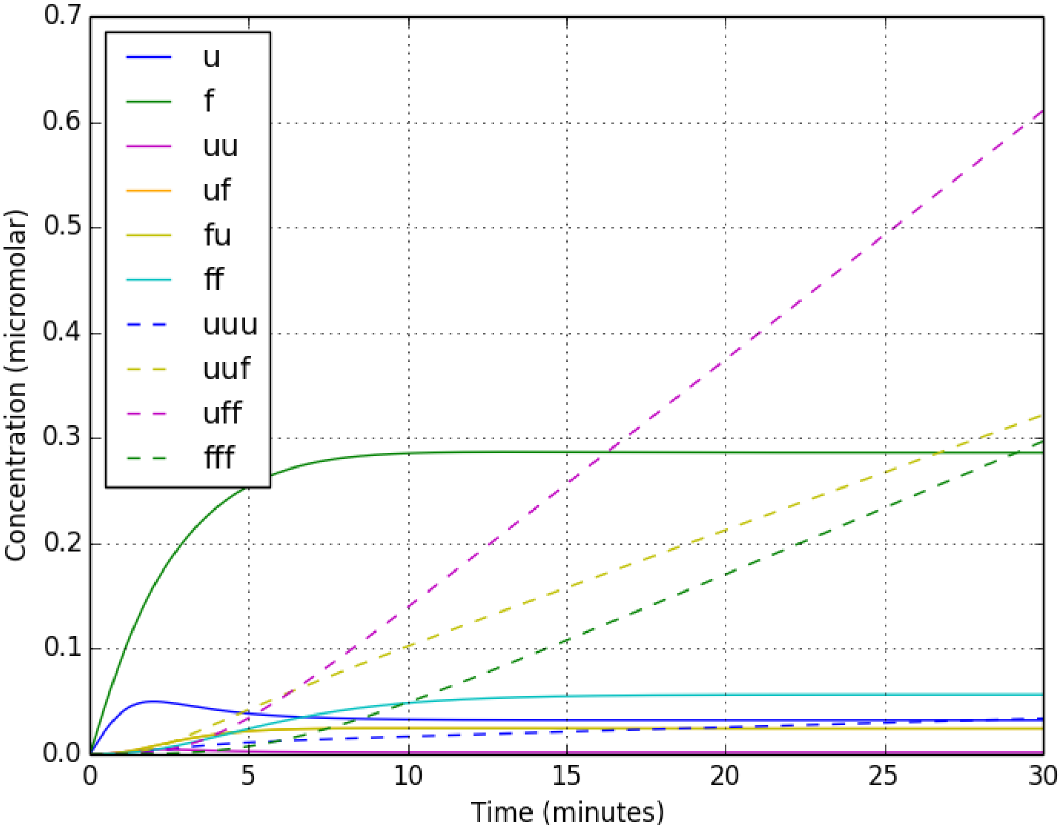


FIG. 8: The change in concentration of our 10 species over time, produced by a python script that solves our set of 10 coupled nonlinear differential equations over 30 minutes. After about 15 minutes, the uff trimer is the major product. This model estimates that after 25 minutes of protein production, about 300 molecules of the uff trimer have been produced in a single *E. coli* cell.

Method 2: Gillespie algorithm

The Gillespie method is a stochastic simulation algorithm that accounts for two assumptions in the reaction rate equation method. First, the reaction rate equations treat the number of molecules as a continuous quantity rather than a discrete quantity. Second, the reaction rate equations are deterministic, though molecular interactions are not truly deterministic but rather probabilistic. While the effects of these assumptions are inconsequential when dealing with a large number of molecules, they become significant when dealing with a small number of molecules, as we are.

We can use the Gillespie Stochastic Simulation Algorithm with the assumptions that the molecules are sufficiently complex so that the effects of quantum mechanics are negligible (the molecules mostly follow Newton's laws of motion), and that the solution is well mixed (meaning that non-reactive collisions occur more frequently than reactive ones), so the fast dynamics of the system can be neglected and the system can be represented simply by the number of molecules in it [14]. We can turn our reaction rate equations into discrete, probabilistic equations for use in the Gillespie method; these are called propensity functions. The probability that a certain reaction j occurs within time dt is $a_j$dt, where a_j is the propensity function for the reaction. The propensity function for a reaction between A and B to form AB is

$$a_{AB} = c_{AB} R_A R_B$$

where R_A and R_B are the number of molecules of each reactant, and the c_{AB} is a rate constant derived from the old rate constant k_{AB} as

$$c_{AB} = k_{AB}/(N_A V) \tag{21}$$

[15]. Here, V is the volume of an *E. coli* cell (about 1 cubic micrometer, or 1×10^{-15} liter), and N_A is Avogadro's number (to convert the dimensions from moles^{-1}s^{-1} to molecules^{-1}s^{-1}). However, if the two species being counted are identical, then

$$a_{AA} = c_{AA} R_A (R_A - 1)/2$$

where $R_A(R_A - 1)/2$ is the number of distinct pairs of molecule A, and

$$c_{AA} = 2k_{AA}/(N_A V)$$

[15]. We apply this to our situation as follows. The state of our system is recorded with the column vector, $\mathbf{X}(t)$, which contains the number of molecules of each species at a given

time:

$$\mathbf{X}(t) = \begin{bmatrix} X_0(t) \\ X_1(t) \\ X_2(t) \\ X_3(t) \\ X_4(t) \\ X_5(t) \\ X_6(t) \\ X_7(t) \\ X_8(t) \\ X_9(t) \end{bmatrix} = \begin{bmatrix} u \\ f \\ uu \\ uf \\ fu \\ ff \\ uuu \\ uuf \\ uff \\ fff \end{bmatrix} \tag{22}$$

We then get fourteen propensity equations, corresponding to our fourteen reactions. Note here that the production rates for production of the fused and unfused monomers from DNA (formerly c_u and c_f) are now q_u and q_f, relabeled so as not to confuse them with our new rate constants and to note dimensional changes; $q_u = c_u N_A V$.

$$a_0 = q_u \tag{23}$$

$$a_1 = q_f \tag{24}$$

$$a_2 = c_{uu} X_0(t)(X_0(t) - 1)/2 \tag{25}$$

$$a_3 = c_{uf} X_1(t) X_0(t) \tag{26}$$

$$a_4 = c_{uf} X_0(t) X_1(t) \tag{27}$$

$$a_5 = c_{ff} X_1(t)(X_1(t) - 1)/2 \tag{28}$$

$$a_6 = c_{uuu} X_0(t) X_2(t) \tag{29}$$

$$a_7 = c_{uuf} X_1(t) X_2(t) \tag{30}$$

$$a_8 = c_{ufu} X_0(t) X_3(t) \tag{31}$$

$$a_9 = c_{ufu} X_0(t) X_4(t) \tag{32}$$

$$a_{10} = c_{fuf} X_1(t) X_3(t) \tag{33}$$

$$a_{11} = c_{fuf} X_1(t) X_4(t) \tag{34}$$

$$a_{12} = c_{uff} X_0(t) X_5(t) \tag{35}$$

$$a_{13} = c_{fff} X_1(t) X_5(t) \tag{36}$$

With the Gillespie method, we assume that only one reaction occurs at a time. The probability of a reaction j occurring is $a_j/\sum_{i=0}^{13} a_i$. To determine which reaction happens next, we cumulatively sum the propensity functions and divide by the total sum of the propensity equations, so that the reaction probabilities are distributed proportionate to the values of their propensity equations over the number line from 0 to 1. This creates a vector that looks like this:

$$a_{cumsum} = \begin{bmatrix} \frac{a_0}{\sum_{i=0}^{13}(a_i)} \\ \frac{a_0 + a_1}{\sum_{i=0}^{13}(a_i)} \\ \frac{a_0 + a_1 + a_2}{\sum_{i=0}^{13}(a_i)} \\ \vdots \end{bmatrix} \tag{37}$$

Next, we take a random number, ξ_1, from a uniform distribution from 0 to 1. The index of the smallest entry of a_{cumsum} that is greater than this number will be the reaction that occurs next. To determine when this reaction will occur, we pick another random number, ξ_2, between 0 and 1 and set the time step (τ) to

$$\tau = \frac{\ln(1/\xi_2)}{\sum_{i=0}^{13}(a_i)} \tag{38}$$

To update our state vector to show that the reaction occurred, we use a reaction matrix, $\mathbf{V}$, in which the rows are species (in the same order as for $\mathbf{X}(t)$) and the columns are reactions, each indicating how the number of molecules of each species changes when that reaction occurs. Starting with column 0, reactions are listed in the same order as their propensity equations are listed on the previous page. If the reaction matrix $\mathbf{V}$ were split up into 14 column vectors, each representing the change of state brought about by a particular reaction, those vectors would form a set of all possible reactions involving the fused and unfused monomers. It is for ease of use and notation that they are represented together as a matrix.

$$\mathbf{V} = \begin{bmatrix} 1 & 0 & -2 & -1 & -1 & 0 & -1 & 0 & -1 & -1 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 & -1 & -2 & 0 & -1 & 0 & 0 & -1 & -1 & 0 & -1 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{39}$$

We add the column from $\mathbf{V}$ that corresponds to the next reaction to $\mathbf{X}(t)$ to update our state at time $t + \tau$. Then, the process repeats. This is condensed into the following steps:

1. Evaluate $a_i(\mathbf{X}(t))_{i=0}^{13}$ and $a_{sum}(\mathbf{X}(t)) := \sum_{i=0}^{13} a_i(\mathbf{X}(t))$.
2. Draw two independent uniform random numbers from 0 to 1, ξ_1 and ξ_2.
3. Set j to be the smallest integer satisfying $\sum_{i=0}^{j} a_i(\mathbf{X}(t)) > a_{sum}(\mathbf{X}(t))$.
4. Set $\tau = \dfrac{\ln(1/\xi_2)}{\sum_{i=0}^{13}(a_i)}$.
5. Set $\mathbf{X}(t+\tau) = \mathbf{X}(t) + V[:, j]$.
6. Return to step 1.

These steps and the previous discussion were paraphrased from the cited Higham paper [16]. For further information about the theory behind the Gillespie method and applications of it, refer to this source as well as references [14] and [15].

We wrote these steps into a python code to plot the number of molecules of each species present over time. In figure 9 we see a sample run of the Gillespie algorithm. As with the deterministic rate equations method, we see that the Gillespie method also predicts the production of about 300 uff trimers after 25 minutes of protein production. To assess how widely the results varied between different runs of the Gillespie algorithm, we ran it multiple times. Four runs are shown simultaneously in figure 10. Though there is variance between the runs, the expected number of uff trimers at 25 minutes is around 300 and uff is consistently the dominant species after about 18 minutes.

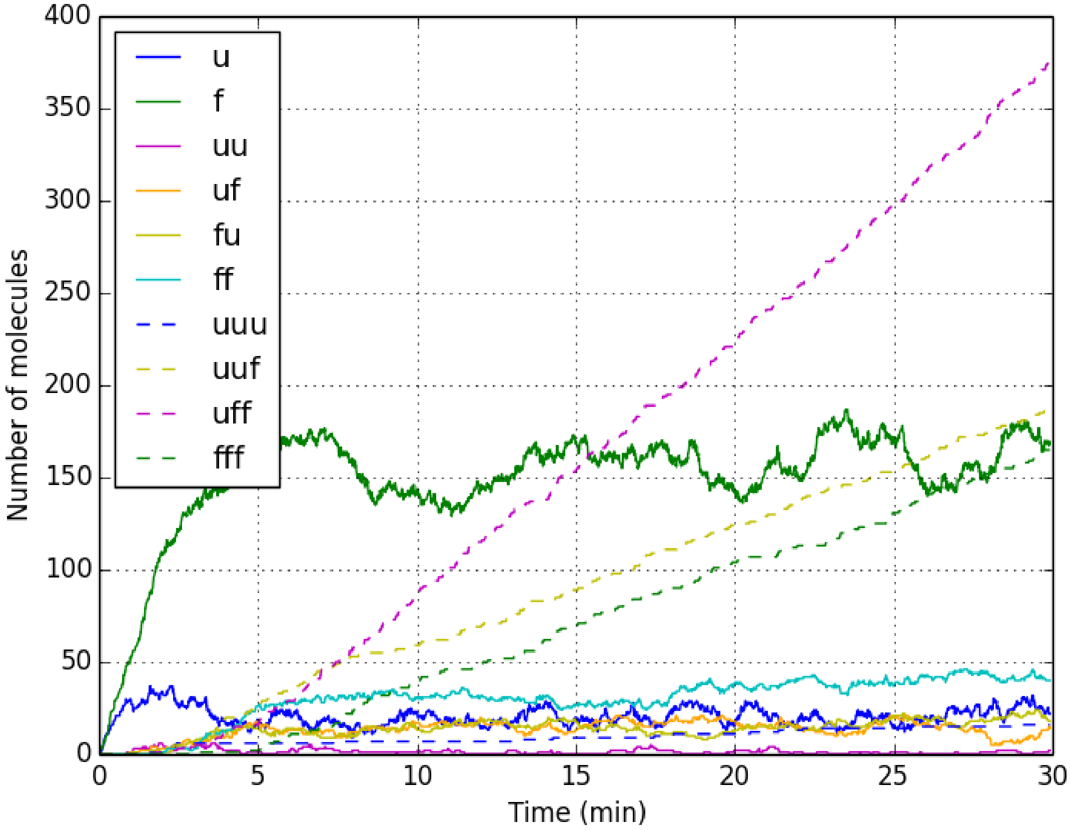


FIG. 9: The change in number of molecules of our 10 species over time, as plotted by the Gillespie algorithm.

DISCUSSION

Our work suggests that making the BSP hexagons as described in this paper is feasible. Unfortunately, it proved challenging to find other research to compare our results to; it appears that number of molecules output by each cell and output rate per cell are not commonly used metrics.

Additionally, the methods made to arrive at approximations here are far from reflecting experimental reality. Many factors that influence protein production rates - such as regulation of gene expression and the extra DNA that is introduced into the genome along with the target gene when using plasmids for gene transfer - were not included here. This is an approximation. The method we used to estimate rate constants for our reactions is overly simplistic and involves a number of broad assumptions. When reviewing these, it is

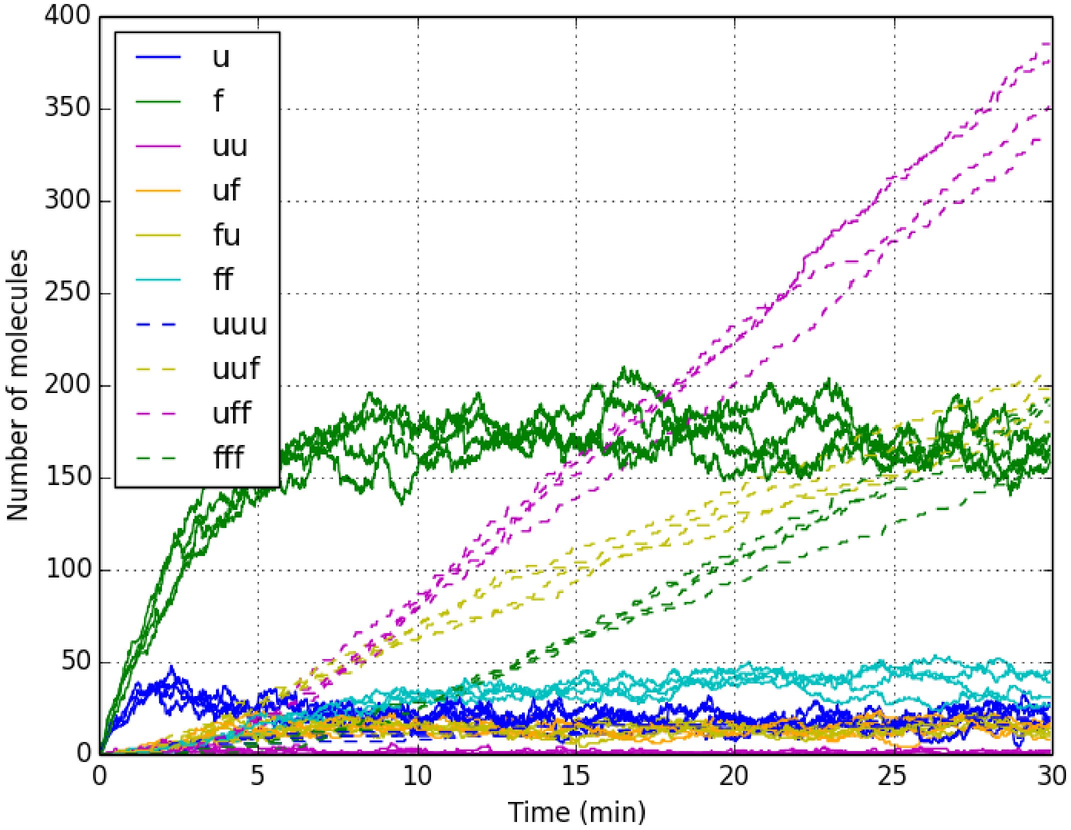


FIG. 10: Four runs of the Gillespie algorithm plotted simultaneously.

important to keep in mind the goals of this project. We do not aim here to form a precise model and explanation of how these reactions would happen in the cell, but rather to obtain a rough estimate of the protein yield we could expect when actually performing this experiment, so as to judge whether performing this experiment would be worthwhile.

Since the results of this project suggest that the BSP hexagons could be created in sufficient yield, the next step will be for our collaborators (the Michael Toney lab in the Department of Chemistry) to carry out the experiment in the lab, guided by our findings here. From our work here, they will know to use a 15:1 gene ratio to achieve the desired 2:1 fused to unfused monomer ratio, and will have estimates for the amounts and timescale of protein production.

ACKNOWLEDGMENTS

I would like to thank Professor Daniel Cox and PhD student Rachel Baarda for their mentorship on this project. Thanks to Krishna Ravikumar, Amanda Parker, and Russell Hawkins for additional advice and feedback on this project. This project was made possible by the Research Experience for Undergraduates program at the UC Davis Department of Physics, led by Professor Rena Zieve. This research was funded by NSF grants numbers PHY-1560482 and DMR-120762 and the UC Davis Office of Research.

Appendix A: 1RFO versus 4NCU

Different papers referring to the T4 bacteriophage fibritin foldon domain use 1RFO and 4NCU as the PDB codes for the protein. Comparing the two proteins from their PDB files in VMD, 4NCU appears to be just one monomer of the foldon domain, whereas 1RFO is the complete trimer. Additionally, 1RFO has 4.5 times as many atoms as 4NCU, whereas one would expect the trimer to have 3 times as many atoms as the monomer. Upon further examination, it appears that the 4NCU protein does not include hydrogen atoms, whereas 1RFO does.

[1] J. Zheng, P. E. Constantinou, C. Micheel, A. P. Alivisatos, R. A. Kiehl, and N. C. Seeman, Nano Lett **6**, 1502 (2006).

[2] N. C. Seeman, Annu Rev Biochem **79**, 65 (2010).

[3] T. Tørring and K. V. Gothelf, F1000Prime Rep **5**, 14 (2013).

[4] H. Gradišar and R. Jerala, Journal of Nanobiotechnology **12** (2014).

[5] E. K. Leinala, P. L. Davies, D. Doucet, M. G. Tyshenko, V. K. Walker, and Z. Jia, J Biol Chem **277**, 33349 (2002).

[6] L. P. Heinz, K. M. Ravikumar, and D. L. Cox, Nano Lett. **15**, 3035 (2015).

[7] M. Peralta, A. Karsai, A. Ngo, C. Sierra, K. Fong, N. Hayre, N. Mirzaee, K. Ravikumar, A. Kluber, X. Chen, G. Liu, M. Toney, R. Singh, and D. Cox, ACS Nano. **9**, 449 (2015).

[8] R. Milo and R. Phillips, *Cell biology by the numbers* (Garland Science, 2015).

[9] S. Güthe, L. Kapinos, A. Möglich, S. Meier, S. Grzesiek, and T. Kiefhaber, J. Mol. Biol. **337**, 905 (2004).

[10] O. G. Berg and P. H. von Hippel, Annu. Rev. Biophys. Biophys. Chem. **14**, 131 (1985).

[11] G. van den Bogaart, *On the mobility of biomolecules: a fluorescence microscopy approach*, Ph.D. thesis, University of Groningen (2008).

[12] I. Y. Torshin, I. T. Weber, and R. W. Harrison, Protein Eng. Des. Sel. **15**, 359 (2002).

[13] Z.-R. Xie, J. Chen, and Y. Wu, J. Phys. Chem. B **120**, 621 (2016).

[14] M. A. Gibson and J. Bruck, J. Phys. Chem. A **104**, 1876 (2000).

[15] F. Hayot and C. Jayaprakash, "A tutorial on cellular stochasticity and gillespies algorithm (draft)," (2006).

[16] D. J. Higham, SIAM Review **50**, 347 (2008).