Fluorescence distributions in combinatorial models of amyloid fibrils composed of split-YFP, Sup35p, and CFP

Daniel Tianming Chen
dchen0@pm.me

May 16, 2026

Abstract

We study exact null-model fluorescence statistics for four combinatorial models of amyloid fibrils: split-YFP alone, split-YFP with Sup35p, FRET YFP/CFP, and FRET YFP/CFP with Sup35p. In their natural units, the split-YFP statistics $R_{\mathrm{SY}}(n) := 2E(n)/n$ and $R_{\mathrm{SY},S}(n) := 2E(n)/n$ (fractions of proteins participating in reconstituted pairs) converge to 2/3 and 1/3, while the FRET statistics $R_{\mathrm{F}}(n) := E(n)/n$ and $R_{\mathrm{F},S}(n) := E(n)/n$ (total adjacent-pair fluorescence per protein) converge to 1/2 and 2/9. These two ratio families measure different physical quantities and should not be read as directly comparable fluorescence yields. For split-YFP systems, standard bivariate generating-function and transfer-matrix methods give exact probability distributions $P(n, k)$, represented either by closed binomial sums or by rational bivariate generating functions, and closed-form evaluations of $k \cdot 2^k$-weighted sums along the shallow diagonals of Pascal's triangle—identities related to known On-Line Encyclopedia of Integer Sequences (OEIS) entries A127976 and A095977, but arising here in a new biophysical context. The same framework unifies the Sup35p sections: in each case the joint distribution $P(n, k)$ is encoded by a rational generating function of the form $(1 + \alpha(w)z)/(1 - \beta(w)z - \gamma(w)z^2)$, from which expected values and variances follow by differentiation at $w = 1$. These results are independently verified by Markov chain arguments and (in the unequal-concentration setting) by Monte Carlo spot checks. For FRET systems, linearity of expectation also gives a finite interaction-radius baseline, with $E_r(n)/n \to 2rp_Cp_Y$ under an additive unit-weight cutoff. All expectation and variance formulas are exact for finite n, not merely asymptotic.

1 Introduction

Amyloids are fibrillar protein aggregates characterized by their distinctive β-sheet structure, where proteins self-assemble into highly ordered stacks [1]. For the purposes of our mathematical analysis, we abstract these complex biochemical structures into a simplified model where individual proteins, denoted A,B,C,..., form sequential patterns in their stacking arrangements (e.g., ABBACCBCBCA...). This abstraction turns a biophysical setting into a tractable combinatorial null model rather than a mechanistic description of fibril growth.

Our combinatorial approach follows the general framework of analytic combinatorics [2, Ch. I]. We study amyloids of different compositions: split Yellow Fluorescent Proteins (split-YFP), a technique based on bimolecular fluorescence complementation [5, 6]; split-YFP together with *Saccharomyces cerevisiae* eukaryotic translation release factor Sup35p,

a prion-forming protein [13]; and YFP/Cyan Fluorescent Protein (CFP) systems under donor excitation that can produce Förster resonance energy transfer (FRET) [3, 14]. A specific quantity of interest is the expected number of *fluorescing locations*—points between two adjacent proteins in the amyloid stack at which a fluorescent interaction occurs (a fusion in the split-YFP settings, a $\{C, Y\}$ adjacency in the FRET settings). We derive the full probability distributions $P(n, k)$ for the number of fluorescing sites and closed-form expressions for the expected values using standard generating functions, transfer matrices, Markov chains, and regular expressions. The symbol $\mathcal{A}(n, k)$ is used locally within each model for the number of length-n amyloids with k fluorescing sites.

In the split-YFP settings, the adjacency blocking constraint leads to weighted shallow-diagonal sums such as $\sum_{k=1}^{\lfloor n/2 \rfloor} k\, 2^k \binom{n-k}{k}$, which we evaluate via generating function techniques [17]. Direct summation and Markov chain arguments yield the same closed forms, providing mutual verification. The closed-form identities that arise turn out to match known OEIS sequences A127976 and A095977 [12, 11]; the contribution of this work is not the identities themselves, but the exact null-model statistics obtained by applying this familiar combinatorial pipeline to a family of related fluorescence models.

Several modeling assumptions should be stated at the outset. First, proteins are assumed to be independently and uniformly distributed in the amyloid sequence. Real amyloid assembly is templated and cooperative [1], so this i.i.d. assumption is a baseline against which non-random assembly effects could be detected. Second, Sup35p is used here purely as an abstract inert third species: the model does not attempt to describe Sup35p's own prion biology, only the effect of inserting a non-fluorescent stack partner. Third, in the split-YFP sections, we model fluorophore reconstitution as a local pairing rule; the rule is motivated by bimolecular fluorescence complementation being effectively irreversible on experimental timescales [5, 6], but abstracts away geometric and kinetic details of real fibril growth. Fourth, in the FRET sections, we initially restrict interactions to nearest neighbors in the fibril, then record the expected-value extension to a finite interaction radius in Section 6. In practice, the inter-strand spacing in cross-β amyloid structure is approximately 0.47 nm [16], while reported Förster radii for common CFP–YFP variants are on the order of 5 nm [14, 7], so the Förster scale is on the order of ten amyloid spacings. The effective interaction along a fibril nevertheless depends on fluorophore geometry and orientation and decays steeply with distance, so this scale should not be read as a hard interaction radius. The nearest-neighbor restriction is therefore a deliberately simplified $r = 1$ baseline, not the quantitative model one would compare directly to experiment. Finally, the equal-concentration assumption ($p = 1/2$ or $1/3$ per protein type) is adopted throughout the main sections; limiting unequal-concentration formulas are discussed in Section 7.

Throughout, K_n denotes the fluorescence count in a length-n amyloid under the model currently being discussed, $E(n) = \mathbb{E}[K_n]$, and $P(n, k) = \mathbb{P}(K_n = k)$. The counting symbol $\mathcal{A}(n, k)$ is model-local: it denotes the number of length-n amyloids with $K_n = k$ in the current section, with the empty word convention $\mathcal{A}(0, 0) = 1$ when bivariate generating functions are used. In split-YFP models, K_n counts fused A–B pairs, so $2K_n$ proteins participate in fluorescence. In FRET models, K_n counts additive C–Y interaction units; a single YFP adjacent to two CFPs contributes two units, so K_n is not generally the number of distinct fluorescent YFP molecules.

Model	Alphabet	Interaction	Native ratio
split-YFP	$\{A, B\}$	A–B fusion	$R_{\mathrm{SY}} = 2E/n$
split-YFP + Sup35p	$\{A, B, S\}$	A–B fusion	$R_{\mathrm{SY},S} = 2E/n$
FRET YFP/CFP	$\{C, Y\}$	C–Y adjacency	$R_{\mathrm{F}} = E/n$
FRET + Sup35p	$\{C, Y, S\}$	C–Y adjacency	$R_{\mathrm{F},S} = E/n$

The paper is organized as follows. Sections 2 and 3 treat split-YFP models, where the blocking constraint requires constrained-string enumeration and finite-state Markov chains. Sections 4 and 5 treat FRET models, where the absence of blocking makes the expectation a direct adjacent-pair count, although transfer matrices still encode the full distribution. Section 6 records finite-n variances and the finite-radius FRET expectation, and Section 7 gives limiting unequal-concentration formulas.

2 Analysis of exclusively split-YFP Amyloids

Let us first analyze amyloids consisting solely of YFP. YFP is a single polypeptide chain that can be split into two fragments, referred to as split-YFP. Call the distinct halves A and B. It is when these two halves A and B fuse that fluorescence occurs at the point of contact. There are no restrictions on the patterns in the amyloid - one can view all combinations of such halves via the regular expression $\{A, B\}^*$.

We are interested in positions of an amyloid where A and B come together and fuse, producing yellow fluorescence. Such a scenario is essentially the reconstitution of the full fluorophore structure of YFP. We refer to such points in the amyloid stack as *fluorescing locations* (or *fluorescing sites*) throughout this paper. The same terminology applies in Sections 4–5 to the FRET settings, where a fluorescing location is a $\{C, Y\}$ adjacency at which energy transfer occurs.

Let n be the number of split-YFPs in a given amyloid. Clearly, there are 2^n possible amyloids of length n.

We will start with the case where the halves of Split-YFP are well-mixed in a neutral solution and of equal concentrations, and are thus equally likely to appear as the next protein in an amyloid sequence.

2.1 Lighting Patterns and 01 representations

We abstract away from A and B specifically and instead analyze the "lighting patterns" of amyloids of length n, defined as follows.

We ask: How many distinct patterns of fluorescing locations could we possibly see from an amyloid of length n?

We define the *01 representation* (or *01 pattern*) of an amyloid $s_1 s_2 \cdots s_n \in \{A, B\}^n$ as the binary string $f_1 f_2 \cdots f_n \in \{0, 1\}^n$ given by:

$$f_i = \begin{cases} 0, & \text{if } i = 1, \\ 1, & \text{if } i \geq 2,\ s_i \neq s_{i-1},\ \text{and } f_{i-1} = 0, \\ 0, & \text{otherwise.} \end{cases}$$

Informally, $f_i = 1$ indicates that a fusion occurs between positions $i - 1$ and i, attributed to the later index i: the protein s_i differs from its neighbor s_{i-1} and that neighbor was not

already consumed by a prior fusion. The first protein always receives $f_1 = 0$ since there is no predecessor. The blocking constraint ($f_{i-1} = 0$ required) ensures that both proteins in a fused pair are "used up," preventing cascading fusions. Note that the 01 representation has the same length as the amyloid, and the mapping f is well-defined for any sequence in $\{A, B\}^n$.

The left-to-right, greedy pairing rule reflects an idealized sequential-polymerization picture: proteins are incorporated one at a time at a single growing end of the fibril, and each newly added protein either fuses with the current endpoint or does not. This greedy convention is motivated by the kinetics of sequential assembly—once a split-YFP pair reconstitutes, bimolecular fluorescence complementation is commonly treated as effectively irreversible on experimental timescales [5, 6], so later arrivals cannot displace it in this null model. It should nevertheless be read as a modeling convention, not as a full kinetic model of fibril growth at multiple ends or with spatial rearrangements.

The convention is also compatible with a common static alternative at the level of counts. If one forms the path graph whose edges are adjacent unlike pairs and asks only for a maximum-cardinality matching, the leftmost greedy algorithm gives a maximum matching on a path: after taking the leftmost available edge, no maximum matching can use more than one edge among its two endpoints, and the remaining problem is again a shorter path. Thus the fluorescence count here agrees with the static maximum-matching count under unit weights, although the identity of the paired proteins remains tied to the chosen growth convention. Other static ensembles, such as Boltzmann-weighted matchings or geometry-dependent pairing propensities, would define different null models and are outside the scope of the present calculation.

For example, if we have ABBAAAAABA, then reading from left to right (because we have to think about how these proteins were added one at a time to analyze which ones fused), we get the 01 representation 0101000010. Writing these two representations side by side shows it:

$$\begin{array}{c} ABBAAAAABA \\ 0\ \ 10\ \ 10\ \ 00\ \ 01\ \ 0 \end{array}$$

We always start with a 0, since the first protein cannot cause the (non-existent) previous protein to fluoresce. Doing this, we no longer need think about whether we started with an A or a B.

As another example,

$$\begin{array}{c} BBBBABABABABABABABAAAABAABABBBBABA \\ 0\ \ 00\ \ 01\ \ 01\ \ 01\ \ 01\ \ 01\ \ 01\ \ 01\ \ 01\ \ 00\ \ 01\ \ 00\ \ 10\ \ 10\ \ 00\ \ 10\ \ 1 \end{array}$$

We observe that an unambiguous regular expression for these 01 patterns is $(0^*01)^*0^*$ (non-ambiguous in the sense required by the symbolic method: each string has a unique parse). Within the mathematical model, if the fluorescing locations were site-resolved, they would determine the 01 pattern: each fluorescing location corresponds to a 1.

We can count the total number of distinct 01 patterns of length n using the symbolic method for regular expressions [2, §I.4]. Note that $(0^*01)^*0^* = (0^+1)^*0^*$: each block 0^*01

consists of zero or more leading zeros followed by 01, so the trailing 0 of 0^* and the leading 0 of 01 merge to give at least one zero before the 1, hence $0^*01 = 0^+1$. We assign weight x per symbol. Each block $0^*01 = 0^+1$ has generating function $\frac{x}{1-x} \cdot x = \frac{x^2}{1-x}$, and the Kleene star $(0^*01)^*$ has generating function $\frac{1}{1-x^2/(1-x)} = \frac{1-x}{1-x-x^2}$. Including the trailing 0^* with generating function $\frac{1}{1-x}$, we obtain:

$$G(x) := \sum_{n=0}^{\infty} a_n x^n = \frac{1-x}{1-x-x^2} \cdot \frac{1}{1-x} = \frac{1}{1-x-x^2}$$

Since $(1-x-x^2)\sum_{n\geq 0} a_n x^n = 1$, we read off the recurrence $a_n = a_{n-1} + a_{n-2}$ for $n \geq 2$, with $a_0 = a_1 = 1$.

We see that $a_0 = 1$, $a_1 = 1$, $a_2 = 2$, $a_3 = 3$, $a_4 = 5$, $a_5 = 8$, etc.

So $a_n = F_{n+1}$, where F_n is the Fibonacci sequence with the standard convention $F_1 = F_2 = 1$. The coefficients $\binom{n-k}{k}$ appearing in this sum are precisely the entries along the *shallow diagonals* of Pascal's triangle, and the identity $a_n = F_{n+1}$ is a well-known consequence of summing along these diagonals. The weighted sums $S_1(n)$ and $S_2(n)$ evaluated in Section 2.4 can be viewed as weighted sums along these same diagonals, with each entry scaled by $k \cdot 2^k$.

2.2 Probability of k fluorescing locations in length n amyloid

We now analyze the probability of seeing k fluorescing locations in an amyloid of length n.

2.2.1 Distinct 01 patterns with k 1s

First, let us count the number of 01 patterns of length n with k 1s. Our invariant states that no 1s may be adjacent, and the first digit must be a zero. Since $f_1 = 0$ always, the k ones must be placed among positions $2, 3, \ldots, n$ (i.e., $n-1$ available positions) with no two adjacent. This is equivalent to placing k non-adjacent objects in $n-1$ boxes, giving:

$$N(n,k) := \binom{n-k}{k}, \quad k \leq \lfloor n/2 \rfloor$$

Indeed, the number of ways to place k non-adjacent objects in m boxes is $\binom{m-k+1}{k}$ (see [15, Exercise 1.34]); setting $m = n-1$ gives $\binom{n-k}{k}$ as claimed.

2.2.2 Distinct amyloids with k 1s

Now, for $k \geq 1$, the number of distinct amyloids of length n with k fluorescing locations is:

$$\mathcal{A}(n,k) = 2^k \cdot N(n-2, k-1) + 2^{k+1} \cdot N(n-1, k) = 2^k \binom{n-k-1}{k-1} + 2^{k+1} \binom{n-k-1}{k}$$

The zero-fluorescence case is $\mathcal{A}(n,0) = 2$, corresponding to the two constant amyloids $AA\cdots A$ and $BB\cdots B$. Throughout, binomial coefficients with lower index outside the usual range are interpreted as zero; nevertheless the displayed formula for $\mathcal{A}(n,k)$ is stated for $k \geq 1$, with $k = 0$ handled separately.

To justify the formula, fix an admissible 01 pattern f with k ones. Once s_1 is chosen, every later protein is determined except immediately after a 1 in the 01 pattern: if $f_i = 1$, then s_i is forced to be the opposite of s_{i-1}; if $f_i = 0$ and $f_{i-1} = 0$, then s_i is forced to equal s_{i-1}; if $f_i = 0$ and $f_{i-1} = 1$, then s_i is free because the previous protein has already been consumed by a fusion. Thus a fixed pattern ending in 0 has 2^{k+1} realizing amyloids (the initial choice, plus one free choice immediately after each of the k ones, all of which have successors), while a fixed pattern ending in 1 has 2^k realizing amyloids.

It remains only to count admissible 01 patterns by their final digit. Patterns ending in 0 are obtained by appending a zero to a length-$(n-1)$ admissible pattern with k ones, giving $N(n-1,k)$ choices. Patterns ending in 1 must have $f_{n-1} = 0$, and deleting the final 01 leaves a length-$(n-2)$ admissible pattern with $k-1$ ones, giving $N(n-2,k-1)$ choices. Multiplying by the fixed-pattern counts proves the displayed formula.

For example, when $k = 5$ the two cases have the following degrees of freedom:

01 pattern ends with a 0

Consider the following pattern. We write the number of possible proteins (A or B) at that location below the corresponding location of the pattern.

$$0000100001010010000010000$$
$$2111121111212112111121111$$

Indeed, for each block of 0's, at the beginning of the block, any of either A or B may appear there, since it will not fuse with the previous protein. Then thereafter each protein is uniquely determined by this choice through to the 1 that appears.

Any pattern of this case therefore has 2^{k+1} possible amyloids associated with it. There are $N(n-1,k)$ possible patterns for this case.

01 pattern ends with a 1

Consider the following 01 pattern. We write the number of possible proteins (A or B) at that location below the corresponding location of the 01 pattern.

$$0000100001010010000000001$$
$$2111121111212112111111111$$

As we can see, putting a 1 at the end of the 01 pattern implies only 2^k possible amyloids with the particular pattern. There are $N(n-2,k-1)$ patterns with this property.

Verification. For $n = 4$, $k = 1$: $\mathcal{A}(4,1) = 2^1\binom{2}{0} + 2^2\binom{2}{1} = 2 + 8 = 10$. For $k = 2$: $\mathcal{A}(4,2) = 2^2\binom{1}{1} + 2^3\binom{1}{2} = 4 + 0 = 4$ (the only 01 pattern is 0101, giving amyloids ABAB, BABA, ABBA, BAAB). With $\mathcal{A}(4,0) = 2$ (AAAA and BBBB), we have $2+10+4 = 16 = 2^4$. ✓

Hence the probability of seeing k fluorescing locations in an amyloid of length n is

$$P(n,k) = \frac{\mathcal{A}(n,k)}{2^n}$$

2.3 Expected number of fluorescing locations

This is simply the sum:

$$E(n) = \frac{1}{2^n} \sum_{k=1}^{\lfloor n/2 \rfloor} k \cdot \mathcal{A}(n,k) = \frac{1}{2^n} \sum_{k=1}^{\lfloor n/2 \rfloor} k \left(2^k \binom{n-k-1}{k-1} + 2^{k+1} \binom{n-k-1}{k} \right)$$

by definition of expected value.

2.4 Expected fraction of proteins involved in fluorescence

Since each fluorescing location involves two proteins, the fraction of proteins participating in fluorescence is:

$$R_{\mathrm{SY}}(n) := \frac{2E(n)}{n} = \frac{1}{n \cdot 2^{n-1}} \sum_{k=1}^{\lfloor n/2 \rfloor} k \left(2^k \binom{n-k-1}{k-1} + 2^{k+1} \binom{n-k-1}{k} \right)$$

We claim that

$$\sum_{k=1}^{\lfloor n/2 \rfloor} k 2^k \binom{n-k-1}{k-1} = \frac{2^n(3n+2) + 2(6n-1)(-1)^n}{27} \tag{1}$$

and

$$\sum_{k=1}^{\lfloor n/2 \rfloor} k 2^k \binom{n-k-1}{k} = \frac{2^n(3n-4) - 2(3n-2)(-1)^n}{27} \tag{2}$$

We prove both identities using ordinary generating functions.

We will use the identity (see, e.g., [17, §1.2]) that for any fixed $j \geq 0$,

$$\sum_{n \geq 2j} \binom{n-j}{j} x^n = \frac{x^{2j}}{(1-x)^{j+1}} \tag{3}$$

which follows from the substitution $m = n - j$ and the known generating function

$$\sum_{m \geq j} \binom{m}{j} x^m = \frac{x^j}{(1-x)^{j+1}}.$$

Proof of (1). Let $S_1(n) := \sum_{k=1}^{\lfloor n/2 \rfloor} k\, 2^k \binom{n-k-1}{k-1}$. We derive the ordinary generating function $G_1(x) = \sum_{n \geq 0} S_1(n) x^n$.

Using the substitution $j = k - 1$,

$$S_1(n) = \sum_{j \geq 0} (j+1)\, 2^{j+1} \binom{n-j-2}{j}$$

Working as a formal power series, we may interchange the order of summation termwise. Applying (3) with a shift $n \mapsto n - 2$:

$$G_1(x) = \sum_{j \geq 0} (j+1)\, 2^{j+1} \cdot \frac{x^{2j+2}}{(1-x)^{j+1}} = \frac{2x^2}{1-x} \sum_{j \geq 0} (j+1) \left(\frac{2x^2}{1-x} \right)^j$$

Setting $u = 2x^2/(1-x)$ and using $\sum_{j\geq 0}(j+1)u^j = 1/(1-u)^2$, we have

$$G_1(x) = \frac{2x^2}{1-x} \cdot \frac{1}{(1-u)^2}$$

Since $1 - u = (1 - x - 2x^2)/(1-x) = (1-2x)(1+x)/(1-x)$, this simplifies to

$$G_1(x) = \frac{2x^2(1-x)}{(1-2x)^2(1+x)^2}$$

We now perform a partial fraction decomposition:

$$G_1(x) = \frac{\alpha}{1-2x} + \frac{\beta}{(1-2x)^2} + \frac{\gamma}{1+x} + \frac{\delta}{(1+x)^2}$$

Multiplying both sides by $(1-2x)^2(1+x)^2$ and evaluating at $x = 1/2$ gives $\beta = 1/9$, and at $x = -1$ gives $\delta = 4/9$. Expanding the right-hand side and comparing coefficients of x^0 and x^1 then yields $\alpha = -1/27$ and $\gamma = -14/27$. Now, using the standard power series

$$\frac{1}{1-2x} = \sum_{n\geq 0} 2^n x^n, \qquad \frac{1}{(1-2x)^2} = \sum_{n\geq 0}(n+1)\,2^n x^n$$

$$\frac{1}{1+x} = \sum_{n\geq 0}(-1)^n x^n, \qquad \frac{1}{(1+x)^2} = \sum_{n\geq 0}(n+1)(-1)^n x^n$$

we extract the coefficient of x^n:

$$\begin{aligned} S_1(n) &= -\frac{1}{27}\cdot 2^n + \frac{1}{9}(n+1)\,2^n - \frac{14}{27}(-1)^n + \frac{4}{9}(n+1)(-1)^n \\ &= \frac{1}{27}\left[2^n(-1+3n+3) + (-1)^n(-14+12n+12)\right] \\ &= \frac{2^n(3n+2) + 2(6n-1)(-1)^n}{27} \end{aligned}$$

as claimed. □

Proof of (2). Let $S_2(n) := \sum_{k=1}^{\lfloor n/2\rfloor} k\,2^k\binom{n-k-1}{k}$. As before, applying (3):

$$G_2(x) = \sum_{n\geq 0} S_2(n)x^n = \sum_{k\geq 1} k\cdot 2^k \cdot \frac{x^{2k+1}}{(1-x)^{k+1}} = \frac{x}{1-x}\cdot\frac{u}{(1-u)^2}$$

where $u = 2x^2/(1-x)$ as before. This simplifies to

$$G_2(x) = \frac{2x^3}{(1-2x)^2(1+x)^2}$$

Performing partial fractions, evaluating at $x = 1/2$ gives $\beta = 1/9$ and at $x = -1$ gives $\delta = -2/9$. Comparing coefficients of x^0 and x^1 yields $\alpha = -7/27$ and $\gamma = 10/27$. Extracting coefficients:

$$\begin{aligned} S_2(n) &= -\frac{7}{27}\cdot 2^n + \frac{1}{9}(n+1)\,2^n + \frac{10}{27}(-1)^n - \frac{2}{9}(n+1)(-1)^n \\ &= \frac{1}{27}\left[2^n(-7+3n+3) + (-1)^n(10-6n-6)\right] \\ &= \frac{2^n(3n-4) - 2(3n-2)(-1)^n}{27} \end{aligned}$$

as claimed. □

Substituting (1) and (2):

$$R_{\mathrm{SY}}(n) = \frac{1}{n \cdot 2^{n-1}} \sum_{k=1}^{\lfloor n/2 \rfloor} k \left(2^k \binom{n-k-1}{k-1} + 2^{k+1} \binom{n-k-1}{k} \right) \tag{4}$$

$$= \frac{1}{n \cdot 2^{n-1}} \left(\frac{2^n(3n+2) + 2(6n-1)(-1)^n}{27} + 2 \cdot \frac{2^n(3n-4) - 2(3n-2)(-1)^n}{27} \right) \tag{5}$$

$$= \frac{1}{n \cdot 2^{n-1}} \cdot \frac{2^n(9n-6) + 6(-1)^n}{27} \tag{6}$$

$$= \frac{2(9n-6)}{27n} + \frac{6(-1)^n}{27n \cdot 2^{n-1}} \tag{7}$$

$$= \frac{2(3n-2)}{9n} + \frac{2(-1)^n}{9n \cdot 2^{n-1}} \tag{8}$$

Then as $n \to \infty$, the second term vanishes and we have

$$\lim_{n\to\infty} R_{\mathrm{SY}}(n) = \lim_{n\to\infty} \frac{2(3n-2)}{9n} = \frac{6}{9} = \frac{2}{3} \tag{9}$$

2.4.1 Weighted shallow diagonal identities

The identities (1) and (2) have a natural interpretation in terms of Pascal's triangle. Recall that the entries along the shallow diagonals of Pascal's triangle are $\binom{n-k}{k}$ for $k = 0, 1, \ldots, \lfloor n/2 \rfloor$, and their sum $\sum_{k=0}^{\lfloor n/2 \rfloor} \binom{n-k}{k} = F_{n+1}$ yields the Fibonacci numbers (Section 2). The identities (1) and (2) provide closed-form evaluations of *weighted* sums along these same diagonals, where each entry is scaled by $k \cdot 2^k$:

$$\sum_{k=1}^{\lfloor n/2 \rfloor} k \cdot 2^k \binom{n-k-1}{k-1} = \frac{2^n(3n+2) + 2(6n-1)(-1)^n}{27} \tag{10}$$

$$\sum_{k=1}^{\lfloor n/2 \rfloor} k \cdot 2^k \binom{n-k-1}{k} = \frac{2^n(3n-4) - 2(3n-2)(-1)^n}{27} \tag{11}$$

These identities arise here from the fluorescence analysis, but the sequences themselves are not new. Specifically, $S_1(n) = 2 \cdot \mathrm{A127976}(n-1)$ and $S_2(n) = \mathrm{A095977}(n-2)$, where OEIS lists A127976 as a convolution involving Jacobsthal numbers and A095977 as counting 2×2 tiles in tilings of $3 \times (n+1)$ rectangles [12, 11]. Both sequences satisfy the linear recurrence $a_n = 2a_{n-1} + 3a_{n-2} - 4a_{n-3} - 4a_{n-4}$ with different initial conditions, which can be verified from the generating function denominator $(1+x)^2(1-2x)^2 = 1 - 2x - 3x^2 + 4x^3 + 4x^4$. Direct computation confirms: $S_1(2) = 2 = 2 \cdot \mathrm{A127976}(1)$ and $S_2(3) = 2 = \mathrm{A095977}(1)$, with subsequent values matching by the shared recurrence.

The appearance of the same sequences in these different settings is explained by the shared reduction to constrained binary strings and rational generating functions with denominator $(1+x)^2(1-2x)^2$. Thus the OEIS matches are useful checks and interpretations, but the main point here is their role in the fluorescence null model.

As a corollary, the linear combination $S_1(n)+2\,S_2(n) = \left[2^n(9n-6)+6(-1)^n\right]/27$ governs the expected fluorescence in split-YFP amyloids, connecting these combinatorial identities directly to the biophysical application.

2.4.2 Alternative derivation via Markov chain

We can also derive $E(n)$ directly by a Markov chain argument [8], bypassing $\mathcal{A}(n,k)$ entirely. Consider building an amyloid one protein at a time, where each protein is independently A or B with equal probability 1/2. Define two states based on the rightmost protein:

- **State** p: the rightmost protein is *not* part of a fused pair.
- **State** q: the rightmost protein is part of a fused pair.

From state p (say the rightmost protein is A, unfused): with probability 1/2 the next protein is A (same type, no fusion, stay in p), and with probability 1/2 it is B (fusion, go to q, +1 fluorescence). By symmetry, the case for B is identical. So from p: transition to p with probability 1/2 and to q with probability 1/2, with expected fluorescence 1/2.

From state q (rightmost protein is fused): the next protein is A or B each with probability 1/2, and in either case goes to state p with no fluorescence (since the current protein is already fused). So from q: transition to p with probability 1.

With $p_1 = 1$ (the first protein is always unfused), the recurrence is

$$p_{n+1} = \underbrace{\tfrac{1}{2}\,p_n}_{\text{stay in } p} + \underbrace{(1-p_n)}_{\text{from } q \text{ to } p} = 1 - \tfrac{1}{2}\,p_n$$

which has fixed point $p^* = 2/3$ and general solution

$$p_n = \frac{2}{3} + \frac{1}{3}\left(-\frac{1}{2}\right)^{n-1}$$

The expected fluorescence contributed by the $(n+1)$th protein is $e_{n+1} = \frac{1}{2}p_n$, so

$$\begin{aligned} E(n) = \sum_{i=2}^{n} e_i = \frac{1}{2}\sum_{j=1}^{n-1} p_j &= \frac{1}{2}\sum_{j=1}^{n-1}\left[\frac{2}{3} + \frac{1}{3}\left(-\frac{1}{2}\right)^{j-1}\right] \\ &= \frac{n-1}{3} + \frac{1}{6}\cdot\frac{1-(-1/2)^{n-1}}{3/2} = \frac{n-1}{3} + \frac{1}{9}\left(1 - \left(-\frac{1}{2}\right)^{n-1}\right) \end{aligned}$$

Therefore

$$R_{\mathrm{SY}}(n) = \frac{2E(n)}{n} = \frac{2(n-1)}{3n} + \frac{2}{9n}\left(1 - \left(-\frac{1}{2}\right)^{n-1}\right) \longrightarrow \frac{2}{3} \text{ as } n \to \infty$$

Since this must agree with the combinatorial derivation, equating the two expressions for $E(n)$ provides an independent verification of the sum identities (1) and (2). Specifically,

$$\frac{S_1(n) + 2\,S_2(n)}{2^n} = \frac{n-1}{3} + \frac{1}{9}\left(1 - \left(-\frac{1}{2}\right)^{n-1}\right)$$

To verify this, substitute the closed forms from (1) and (2):

$$\frac{S_1(n) + 2S_2(n)}{2^n} = \frac{2^n(9n-6) + 6(-1)^n}{27 \cdot 2^n} = \frac{3n-2}{9} + \frac{2(-1)^n}{9 \cdot 2^n}$$

while the Markov chain gives $E(n) = \frac{n-1}{3} + \frac{1}{9} - \frac{(-1)^{n-1}}{9 \cdot 2^{n-1}} = \frac{3n-2}{9} + \frac{2(-1)^n}{9 \cdot 2^n}$, confirming the two expressions are identical.

We can also verify the above via computer simulation (see `split_yfp_simulation.py`; all scripts are available at `https://github.com/danielchen0/amyloids`). A sample of computed values is shown in Table 1, and the first 400 values appear in Figure 1.

n	$R_{\mathrm{SY}}(n)$	n	$R_{\mathrm{SY}}(n)$	n	$R_{\mathrm{SY}}(n)$
2	0.500000	10	0.622266	20	0.644444
3	0.500000	11	0.626243	25	0.648889
4	0.562500	12	0.629639	50	0.657778
5	0.575000	13	0.632474	100	0.662222
6	0.593750	14	0.634923	200	0.664444
7	0.602679	15	0.637036	400	0.665556

$R_{\mathrm{SY}}(n) \to 2/3 \approx 0.666667$

Table 1: Sample values of $R_{\mathrm{SY}}(n)$ for split-YFP amyloids.

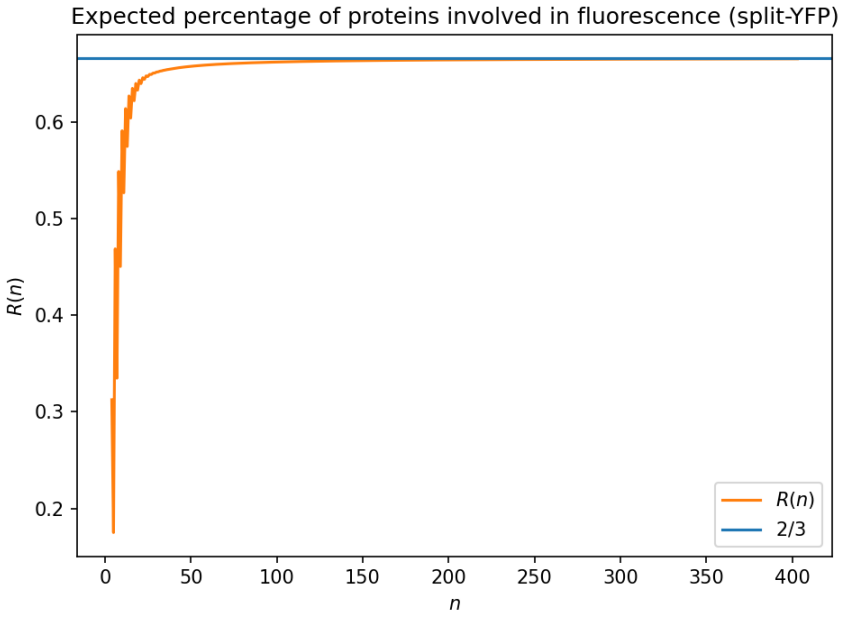

Figure 1: $R_{\mathrm{SY}}(n)$ for split-YFP amyloids. The blue line is 2/3 and the orange curve is the exact $R_{\mathrm{SY}}(n)$. Convergence to the limiting ratio is algebraic ($O(1/n)$), with a smaller alternating exponential correction.

3 Introducing Sup35p into Amyloids

We now want to consider the facts above but with the introduction of yeast prion protein Sup35p (which we label with an S) into our amyloids [13]. In this null model, Sup35p is used as an abstract inert third species: we do not model its own prion biology, only its role as a non-fluorescent stack partner. It cannot participate in any fluorescent interaction, and its sole effect in the model is to sterically separate adjacent split-YFP halves, preventing them from conjoining and fluorescing. Our amyloids can now be described using the regular expression $\{A, B, S\}^*$. We would like to first analyze the case where Sup35p and the halves of Split-YFP come in equal concentrations, and are thus equally likely to occur as the next protein in an amyloid sequence.

The 01 representation introduced in Section 2 must be extended to $\{A, B, S\}^n$ to reflect Sup35p's inertness. Applied verbatim, the Section 2 rule "$f_i = 1$ iff $s_i \neq s_{i-1}$ and $f_{i-1} = 0$" would mark an $S \to A$ adjacency as fluorescing, contradicting the modeling assumption that S blocks fluorescence. The intended rule, used implicitly throughout this section, is:

$$f_i = \begin{cases} 0, & \text{if } i = 1, \\ 1, & \text{if } i \geq 2,\ \{s_{i-1}, s_i\} = \{A, B\},\ \text{and } f_{i-1} = 0, \\ 0, & \text{otherwise.} \end{cases}$$

That is, a 1 requires both proteins in the adjacency to be split-YFP halves (one A and one B), and that the left protein not already be consumed by a prior fusion. With this extension, the set of 01 patterns produced by $\{A, B, S\}^n$ sequences is exactly the same as in Section 2: the regular expression $(0^*01)^*0^*$ still describes all attainable patterns, since insertions of S only stretch existing zero runs.

3.1 Number of distinct lighting patterns and 01 representations

The introduction of Sup35p does not change this. When we abstract away the details we are still looking at 01 patterns of the form $(0^*01)^*0^*$. Hence we have F_{n+1} distinct lighting patterns for an amyloid of length n (using the convention $F_1 = F_2 = 1$).

3.2 Probability of k fluorescing locations in length n amyloid

We derive the joint distribution of length and fluorescence count from a bivariate generating function obtained by a two-state Markov decomposition, and then unpack it into a block-decomposition formula that mirrors Section 2.

Bivariate generating function. Throughout the transfer-matrix derivations, $M(w)_{ij}$ weights a step from state i to state j, row vectors multiply on the left, and a final column vector of ones sums over terminal states. Build an A-B-S amyloid one protein at a time, classifying its rightmost protein into two lumped states:

- p: rightmost is A or B and is not currently part of a fused pair.
- q: rightmost is either S, or is A/B that has just been fused with its left neighbor.

The transitions, weighted by z for length and w for each fluorescing site produced, are:

$$M(w) = \begin{pmatrix} 1 & 1+w \\ 2 & 1 \end{pmatrix},$$

where the rows index the from-state (p, q) and the columns the to-state. From p (say rightmost is A, unfused): next protein A yields one new amyloid ending in p (weight 1); next protein B fuses and yields one amyloid ending in q (weight w, +1 fluorescence); next protein S yields one amyloid ending in q (weight 1). From q: either A or B sends us to p (weight 2 for two letters), and S sends us back to q (weight 1). This two-state lumping is valid here because A and B have equal weights and symmetric transition behavior; for unequal A/B concentrations the unfused A and unfused B states must be separated, as in Section 7. For $n \geq 1$, the row vector of length-n weighted counts is $(2, 1)M(w)^{n-1}$; the initial vector $(2, 1)$ records the two one-letter amyloids A/B in state p and the one one-letter amyloid S in state q. Hence the joint generating function

$$F(z, w) := \sum_{n \geq 0} \sum_{k \geq 0} \mathcal{A}(n, k)\, w^k z^n$$

(where in this section $\mathcal{A}(n, k)$ counts amyloids over $\{A, B, S\}$, and $\mathcal{A}(0, 0) := 1$ by convention) is

$$F = 1 + \sum_{n \geq 1} z^n (2, 1) M(w)^{n-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 1 + z(2, 1)(I - zM(w))^{-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

which evaluates to

$$F(z, w) = \frac{1 + z}{1 - 2z - (1 + 2w)\, z^2}. \tag{12}$$

Equivalently, the coefficients satisfy the recurrence $\mathcal{A}(n, k) = 2\mathcal{A}(n-1, k) + \mathcal{A}(n-2, k) + 2\mathcal{A}(n-2, k-1)$ with the appropriate small-n initial conditions. Two specializations recover familiar facts. At $w = 0$, we count amyloids with no fluorescence: $F(z, 0) = (1 + z)/(1 - 2z - z^2)$, whose coefficients b_n enumerate A-B-S sequences with all-zero 01 representation. At $w = 1$, all amyloids are counted: $F(z, 1) = (1 + z)/(1 - 2z - 3z^2) = 1/(1 - 3z)$ (after canceling the factor $1 + z$ from $1 - 2z - 3z^2 = (1 - 3z)(1 + z)$), so $[z^n]F(z, 1) = 3^n$ as required.

The expected value $E(n)$ can now be extracted in a few lines from $\partial F/\partial w\big|_{w=1}$:

$$\left.\frac{\partial F}{\partial w}\right|_{w=1} = \frac{2z^2(1 + z)}{(1 - 2z - 3z^2)^2} = \frac{2z^2}{(1 - 3z)^2(1 + z)},$$

and partial-fraction expansion (carried out explicitly in Section 3.4 below as a consistency check) yields the closed form

$$E(n) \cdot 3^n = [z^n] \frac{2z^2}{(1 - 3z)^2(1 + z)} = \frac{3^{n-1}(4n - 3) + (-1)^n}{8}.$$

Equivalently, $E(n) = (4n - 3)/24 + (-1)^n/(8 \cdot 3^n) = (4n - 3)/24 - (-1)^{n-1}/(24 \cdot 3^{n-1})$, which is the same formula derived by the explicit Markov recurrence in Section 3.4.

Block decomposition. The bivariate generating function (12) admits a more explicit reading via the regular-expression decomposition $(0^*01)^*0^*$ of the 01 pattern. Each amyloid of length n with k ones in its 01 pattern decomposes uniquely as k consecutive "active blocks" of the form 0^+1, followed by a trailing run of zeros:

$$\underbrace{0\cdots01}_{\phi_1}\ \underbrace{0\cdots01}_{\phi_2}\ \cdots\ \underbrace{0\cdots01}_{\phi_k}\ \underbrace{0\cdots0}_{q}$$

with $\phi_i \geq 2$ and $q \geq 0$. The protein choices within different blocks are independent: once block i ends with a fusion at position j (so s_j is consumed by the previous fusion), the protein at position $j+1$ may be any of $\{A, B, S\}$ unconstrained, since $f_{j+1} = 0$ is forced regardless of s_{j+1}. Hence the number of amyloids realizing a given block decomposition factors:

$$\mathcal{A}(n,k) = \sum_{q=0}^{n-2k} b_q \sum_{\substack{\phi_1+\cdots+\phi_k=n-q \\ \phi_i\geq 2}} \prod_{i=1}^{k} a_{\phi_i}, \tag{13}$$

where a_ϕ counts A-B-S patterns of length ϕ whose 01 representation is $0^{\phi-1}1$, and b_q counts patterns of length q with all-zero 01 representation. For $k = 0$, the formula reduces to the all-zero case $\mathcal{A}(n, 0) = b_n$. Thus $P(n,k) = \mathcal{A}(n,k)/3^n$. Equation (13) is the convolutional unfolding of (12): writing $F(z,w) = B(z)/(1 - w\,A(z))$ with $A(z) = \sum_{\phi\geq 2} a_\phi z^\phi = 2z^2/(1 - 2z - z^2)$ and $B(z) = \sum_{q\geq 0} b_q z^q = (1+z)/(1-2z-z^2)$, one verifies

$$\frac{B(z)}{1 - wA(z)} = \frac{(1+z)/(1-2z-z^2)}{1 - 2wz^2/(1-2z-z^2)} = \frac{1+z}{1-2z-(1+2w)z^2},$$

recovering (12).

Unlike the two-letter split-YFP case, this gives $P(n,k)$ most naturally as a coefficient or convolution formula rather than as a single binomial closed form; the rational bivariate generating function is the compact representation of the full distribution.

The sequences a_n and b_n. Both sequences are governed by the same denominator $1 - 2x - x^2$. From the regular expressions $S^*\{AA^*SS^*, BB^*SS^*\}^*\{AA^*, BB^*, \varepsilon\}$ for the b-sequence and $S^*\{AA^*SS^*, BB^*SS^*\}^*\{AA^*B, BB^*A\}$ for the a-sequence (setting $A = B = S = x$ throughout), one obtains the ordinary generating functions

$$\sum_{n\geq 0} b_n x^n = \frac{1+x}{1-2x-x^2}, \qquad \sum_{n\geq 0} a_n x^n = \frac{2x^2}{1-2x-x^2},$$

yielding the recurrences $b_n = 2b_{n-1} + b_{n-2}$ for $n \geq 2$ (with $b_0 = 1$, $b_1 = 3$) and $a_n = 2a_{n-1} + a_{n-2}$ for $n \geq 3$ (with $a_0 = a_1 = 0$, $a_2 = 2$). The denominator's roots are reciprocals of $1 \pm \sqrt{2}$, and clean partial-fraction decomposition gives the closed forms

$$b_n = \frac{(1+\sqrt{2})^{n+1} + (1-\sqrt{2})^{n+1}}{2}, \quad n \geq 0, \tag{14}$$

$$a_n = \left(1 + \tfrac{\sqrt{2}}{2}\right)(1-\sqrt{2})^n + \left(1 - \tfrac{\sqrt{2}}{2}\right)(1+\sqrt{2})^n, \quad n \geq 1 \tag{15}$$

(the formula for a_n does not extend to $n = 0$, where it gives 2 rather than $a_0 = 0$; this is absorbed by the non-homogeneous initial conditions). Both sequences appear in OEIS: $b_n = \text{A001333}(n+1)$ and $a_n/2 = \text{A000129}(n-1)$ for $n \geq 1$, where A001333 and A000129 are respectively the continued-fraction numerators of $\sqrt{2}$ and the Pell numbers [10, 9]. Verification: $a_2 = (1+\sqrt{2}/2)(3-2\sqrt{2})+(1-\sqrt{2}/2)(3+2\sqrt{2}) = 6-(\sqrt{2}/2)(4\sqrt{2}) = 6-4 = 2$; $b_0 = (1+\sqrt{2}+1-\sqrt{2})/2 = 1$. The first values are $b_0, b_1, \ldots = 1, 3, 7, 17, 41, 99, \ldots$ and $a_0, a_1, \ldots = 0, 0, 2, 4, 10, 24, \ldots$

Worked example. For $n = 4$, $k = 1$, the block decomposition has one block of length $\phi_1 \geq 2$ followed by a tail of length $q = 4 - \phi_1$:

ϕ_1	q	01 pattern	$a_{\phi_1} \cdot b_q$
2	2	0100	$2 \cdot 7 = 14$
3	1	0010	$4 \cdot 3 = 12$
4	0	0001	$10 \cdot 1 = 10$

So $\mathcal{A}(4,1) = 14 + 12 + 10 = 36$, confirmed by brute-force enumeration. Cross-checking against (12): the coefficient of $z^4 w^1$ in $F(z,w)$, computed via the recurrence, equals 36 as well.

3.3 Expected number of fluorescing locations

This is simply the sum:

$$E(n) = \frac{1}{3^n} \sum_{k=1}^{\lfloor n/2 \rfloor} k \cdot \mathcal{A}(n,k)$$

by definition of expected value.

3.4 Expected fraction of proteins involved in fluorescence

This is the number:

$$R_{\text{SY},S}(n) = \frac{2E(n)}{n}$$

We prove that $\lim_{n\to\infty} R_{\text{SY},S}(n) = 1/3$ by computing $E(n)$ directly using a Markov chain argument, bypassing the combinatorial formula for $\mathcal{A}(n,k)$.

Consider building an amyloid one protein at a time, where each protein is independently A, B, or S with equal probability 1/3. We define two states based on the status of the most recently added protein:

- **State** p: the rightmost protein is A or B and is *not* part of a fused pair (and so is available to fuse with the next protein).
- **State** q: the rightmost protein is either S, or is A/B but already part of a fused pair (and so the next protein cannot fuse with it).

The transition probabilities are as follows. From state p (say the rightmost protein is A, unfused):

- With probability 1/3, the next protein is A (same type): no fusion, go to p.

- With probability 1/3, the next protein is B (different type): fusion occurs (+1 fluorescence), go to q.
- With probability 1/3, the next protein is S: no fusion, go to q.

By symmetry between A and B, the case where the rightmost is B is identical. So from p: transition to p with probability 1/3, to q with probability 2/3, with expected fluorescence 1/3.

From state q, the rightmost protein cannot fuse regardless of the next protein's type. The next protein is A, B, or S each with probability 1/3, and is always unfused. So from q: transition to p with probability 2/3 (next is A or B), to q with probability 1/3 (next is S), with expected fluorescence 0.

Let p_n denote the probability of being in state p after the nth protein. The initial state is p with probability 2/3 and q with probability 1/3, so $p_1 = 2/3$. The recurrence is:

$$p_{n+1} = \frac{1}{3}p_n + \frac{2}{3}(1 - p_n) = \frac{2}{3} - \frac{1}{3}p_n$$

This has fixed point $p^* = 1/2$, and the general solution is

$$p_n = \frac{1}{2} + \frac{(-1/3)^{n-1}}{6}$$

which is verified by checking $p_1 = 1/2 + 1/6 = 2/3$.

The expected fluorescence contributed by the $(n+1)$th protein is $e_{n+1} = p_n/3$, so

$$\begin{aligned} E(n) &= \sum_{i=2}^{n} e_i = \frac{1}{3}\sum_{j=1}^{n-1} p_j = \frac{1}{3}\sum_{j=1}^{n-1}\left[\frac{1}{2} + \frac{(-1/3)^{j-1}}{6}\right] \\ &= \frac{n-1}{6} + \frac{1}{18}\cdot\frac{1-(-1/3)^{n-1}}{4/3} = \frac{n-1}{6} + \frac{1}{24}\left(1-(-1/3)^{n-1}\right) \\ &= \frac{4n-3}{24} - \frac{(-1)^{n-1}}{24\cdot 3^{n-1}} \end{aligned}$$

Therefore

$$R_{\mathrm{SY},S}(n) = \frac{2E(n)}{n} = \frac{4n-3}{12n} + \frac{(-1)^n}{12n\cdot 3^{n-1}}$$

and as $n \to \infty$, the second term vanishes:

$$\lim_{n\to\infty} R_{\mathrm{SY},S}(n) = \lim_{n\to\infty}\frac{4n-3}{12n} = \frac{4}{12} = \frac{1}{3}$$

The block decomposition (13) and the transfer-matrix generating function (12) give two views of the same coefficients $\mathcal{A}(n,k)$. Taking expectations then gives the identity

$$\frac{1}{3^n}\sum_{k=1}^{\lfloor n/2\rfloor} k\cdot\mathcal{A}(n,k) = \frac{4n-3}{24} - \frac{(-1)^{n-1}}{24\cdot 3^{n-1}}$$

where $\mathcal{A}(n,k)$ is the combinatorial formula involving the a_{ϕ_i} and b_q sequences with $\sqrt{2}$. The left-hand side is the definition of $E(n)$, and the right-hand side is the closed form

obtained from the transfer matrix or, equivalently, from the explicit Markov recurrence above. This is best read as a consistency check between the coefficient-level distribution and the expectation-level calculation, rather than as two wholly independent Markov-free derivations. Numerically, at $n = 4$, the left-hand side gives $(1 \cdot 36 + 2 \cdot 4)/81 = 44/81$, and the right-hand side gives $(4 \cdot 4 - 3)/24 + 1/(24 \cdot 27) = 13/24 + 1/648 = 44/81$. ✓

A simulation (see `sup35_simulation.py`) confirms this. A sample of computed values is shown in Table 2, and the graph appears in Figure 2.

n	$R_{\mathrm{SY},S}(n)$	n	$R_{\mathrm{SY},S}(n)$	n	$R_{\mathrm{SY},S}(n)$
2	0.222222	10	0.308334	20	0.320833
3	0.246914	11	0.310606	25	0.323333
4	0.271605	12	0.312500	50	0.328333
5	0.283128	13	0.314103	100	0.330833
6	0.291724	14	0.315476	200	0.332083
7	0.297603	15	0.316667	400	0.332708

$R_{\mathrm{SY},S}(n) \to 1/3 \approx 0.333333$

Table 2: Sample values of $R_{\mathrm{SY},S}(n)$ for split-YFP amyloids with Sup35p.

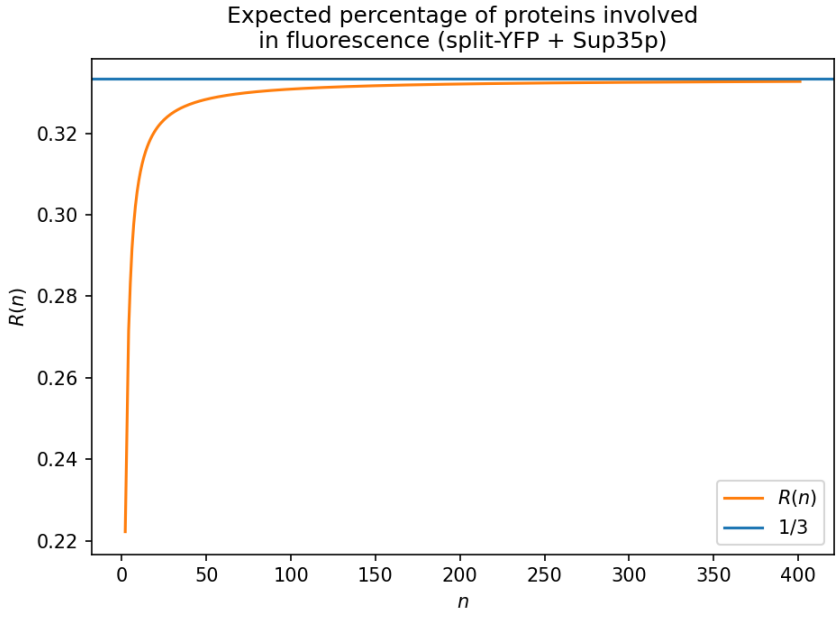

Figure 2: $R_{\mathrm{SY},S}(n)$ for split-YFP amyloids with Sup35p. The blue line is 1/3 and the orange curve is the exact $R_{\mathrm{SY},S}(n)$. Convergence to the limiting ratio is algebraic ($O(1/n)$), with a smaller alternating exponential correction.

4 Using FRET in amyloids of YFP and CFP

Consider the case where intact YFP and CFP are well-mixed in a container and observed under excitation of the CFP donor. When CFP and YFP are in close spatial proximity, nonradiative energy transfer from excited CFP to the YFP acceptor can lead to yellow YFP emission through FRET [3, 14, 7].

We are interested in measuring the *total* yellow fluorescence emanating from amyloids of this kind. In the counting convention used here, an adjacent CFP–YFP pair contributes one unit of yellow fluorescence; a YFP sandwiched between two CFPs therefore contributes two units. This should be read as an additive, time-averaged null-model convention rather than a photophysical prediction about a single acceptor molecule. In practice, real multi-donor FRET to one acceptor does not add linearly in this simple way: efficiencies depend on distance, orientation, donor competition, and saturation effects. We adopt the unit-additive convention for analytical tractability. A more realistic model incorporating geometry-dependent FRET efficiencies would require replacing the simple adjacent-pair counting argument below with a specified physical contribution model.

Label YFP as Y and CFP as C. The blocking constraint $f_{i-1} = 0$ that was central to Sections 2 and 3 does *not* apply here: by our additive convention, both CFP neighbors of a YFP may contribute to the total count, so consecutive 1s in the 01 representation are allowed. In the FRET sections, the 01 pattern is accordingly defined by

$$f_i = \begin{cases} 0, & \text{if } i = 1, \\ 1, & \text{if } i \geq 2 \text{ and } \{s_{i-1}, s_i\} = \{C, Y\}, \\ 0, & \text{otherwise,} \end{cases}$$

with no constraint on f_{i-1}. The protein words are elements of $\{C, Y\}^n$, while the associated 01 patterns are exactly the binary strings of length n whose first symbol is 0. Indeed, after choosing the first protein, each subsequent bit determines whether the next protein is equal to or different from the previous one: a 1 forces it to differ, and a 0 forces it to match. Thus there are 2^{n-1} possible 01 patterns, and each is realized by exactly two Y-C sequences, according to the choice of the first protein.

The number of 1's in the 01 pattern is exactly the total yellow fluorescence in this additive convention: a 1 appears precisely when an adjacent CFP–YFP pair contributes one unit of FRET fluorescence. For example:

$$\begin{aligned} \{C,Y\}^*\text{: } & CYYYCYCYYCC \\ (0^*01^*1)^*0^*\text{: } & 0\ \ 10\ \ 01\ \ 11\ \ 10\ \ 10 \\ \text{Yellow fluorescence: } & 0\ \ 10\ \ 10\ \ 20\ \ 11\ \ 00 \end{aligned}$$

Note that in the 01 pattern, each 1 is attributed to the right endpoint of a $\{C, Y\}$ adjacency as a bookkeeping convention, not to the physically emitting protein. The key claim is that the *total* number of 1s equals the total yellow fluorescence, which we verify: both rows sum to 6.

4.1 Probability of k total yellow fluorescence in length n amyloid

The number of amyloids of length n with k total yellow fluorescence is:

$$\mathcal{A}(n,k) = 2\binom{n-1}{k}, \quad k \in [0, n-1]$$

and thus the probability of observing k total yellow fluorescence in a length n amyloid is

$$P(n,k) = \frac{2\binom{n-1}{k}}{2^n} = \frac{\binom{n-1}{k}}{2^{n-1}}, \quad k \in [0, n-1]$$

.

This is the Binomial$(n-1, 1/2)$ distribution, which can also be seen directly. Let $\varepsilon_i = \pm 1$ according to whether protein i is C or Y, and define $d_i = \varepsilon_i \cdot \varepsilon_{i-1}$ for $i \geq 2$. Then $f_i = 1$ iff $d_i = -1$. The map $(\varepsilon_1, \ldots, \varepsilon_n) \mapsto (\varepsilon_1, d_2, \ldots, d_n)$ is a bijection on $\{\pm 1\}^n$ (since $\varepsilon_i = \varepsilon_1 \prod_{j=2}^{i} d_j$ is recoverable), so $(d_2, \ldots, d_n)$ are mutually independent and uniform on $\{\pm 1\}$. Hence $f_2, \ldots, f_n$ are mutually independent Bernoulli$(1/2)$ despite their overlapping definition, and their sum is Binomial$(n-1, 1/2)$.

4.2 Expected total yellow fluorescence

This is simply the sum:

$$E(n) = \sum_{k=0}^{n-1} k \cdot P(n,k)$$

by definition of expected value.

Note that, by a standard identity [4, §5.1, Eq. 5.5]:

$$E(n) := \frac{1}{2^{n-1}} \cdot \sum_{k=0}^{n-1} k \cdot \binom{n-1}{k} = \frac{1}{2^{n-1}} \cdot (n-1)2^{n-2} = \frac{n-1}{2}$$

So that for very large n, we expect to see about as much yellow unit fluorescence as half the length of the amyloid.

Note that $E(n) = (n-1)/2$ also follows directly from linearity of expectation: the total yellow fluorescence equals the number of adjacent $\{C, Y\}$ pairs, and there are $n-1$ adjacent pairs each with probability $P(\{C,Y\}) = 2 \cdot (1/2)^2 = 1/2$ of being a $\{C, Y\}$ pair (since FRET interactions, unlike split-YFP fusion, have no blocking constraint). We define the fluorescence ratio as

$$R_{\mathrm{F}}(n) := E(n)/n = \frac{n-1}{2n} \longrightarrow \frac{1}{2} \text{ as } n \to \infty$$

The first 400 values of $R_{\mathrm{F}}(n)$ appear in Figure 3.

5 Using FRET in amyloids of YFP, CFP, and Sup35p

As with Section 3, we will now analyze the case of introducing Sup35p to the system in Section 4. Thus we have Sup35p, YFP, and CFP well mixed and of equal concentrations in a container under CFP donor excitation. An adjacent CFP–YFP pair contributes one unit

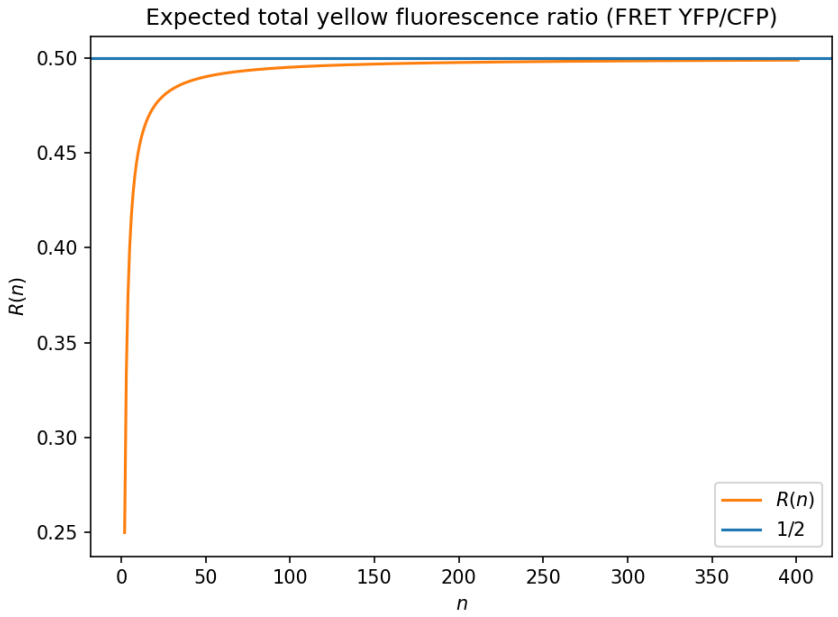

Figure 3: $R_{\mathrm{F}}(n)$ for FRET YFP/CFP amyloids. The blue line is 1/2 and the orange curve is the exact $R_{\mathrm{F}}(n) = (n-1)/(2n)$. Convergence is algebraic ($O(1/n)$), with no exponential correction.

of FRET fluorescence in the additive convention of Section 4; a YFP sandwiched between two CFPs contributes two such units. The presence of Sup35p blocks CFP–YFP adjacency in the one-dimensional model by separating the fluorescent proteins.

Our protein words are elements of $\{C, Y, S\}^*$, where C represents CFP, Y represents YFP, and S represents Sup35p. The 01 representation uses the same local FRET rule as Section 4:

$$f_i = \begin{cases} 0, & \text{if } i = 1, \\ 1, & \text{if } i \geq 2 \text{ and } \{s_{i-1}, s_i\} = \{C, Y\}, \\ 0, & \text{otherwise.} \end{cases}$$

Thus every adjacency involving S contributes 0. As in Section 4, every binary string with first symbol 0 is attainable, since the C-Y-only words already realize all such patterns. Unlike the two-letter case, however, a fixed 01 pattern no longer determines the protein word up to the first letter, because zeros may arise from equal C/Y neighbors or from the presence of S.

5.1 Probability of k total yellow fluorescence in length n amyloid

The block-decomposition approach used for Section 3.2 does not transfer cleanly to this setting. The reason is that, unlike split-YFP fusion, FRET allows consecutive 1s in the 01 pattern (a YFP sandwiched between two CFPs fluoresces twice), so each block has the form 0^+1^+ rather than 0^+1. The last protein of a 1^+ run is always in $\{C, Y\}$, and the first protein of the next block must form $f = 0$ with it—hence must match it or be S, only 2 choices among $\{C, Y, S\}$. By contrast, in Section 3.2, the protein after a 1 is already blocked from

fusing leftward, so the next protein is free among all three letters. This boundary coupling means a FRET block decomposition would need to carry boundary-state information; the transfer matrix is the simpler way to track the rightmost protein type.

Bivariate generating function. Lump $\{C, Y\}$ into a single state (justified by the $C \leftrightarrow Y$ symmetry of the model under equal concentrations), and track two states by the rightmost protein type. Without this symmetry, for example if $p_C \neq p_Y$, this lumping would need to be replaced by a chain tracking C-ending and Y-ending states separately.

- p: rightmost is C or Y.
- q: rightmost is S.

The weighted transition matrix (rows = from, cols = to; z marks length, w marks each fluorescing site) is

$$M(w) = \begin{pmatrix} 1+w & 1 \\ 2 & 1 \end{pmatrix}.$$

From p (say rightmost is C): next protein C stays in p with no fluorescence (weight 1); next protein Y stays in p but produces a fluorescence (weight w); next protein S goes to q (weight 1). By $C \leftrightarrow Y$ symmetry the rightmost-Y case is symmetric. From q (rightmost is S): C and Y carry no fluorescence with S, so next $\in \{C, Y\}$ takes us to p (combined weight 2), and next $= S$ stays in q (weight 1). For $n \geq 1$, the row vector of length-n weighted counts is $(2, 1)M(w)^{n-1}$; again, $(2, 1)$ counts the two possible first letters C/Y in state p and the single first letter S in state q. Thus the joint generating function

$$F_{\mathrm{F},S}(z, w) := \sum_{n \geq 0} \sum_{k \geq 0} \mathcal{A}(n, k)\, w^k z^n$$

(with $\mathcal{A}(0, 0) := 1$) is $1 + z(2, 1)(I - zM(w))^{-1}\binom{1}{1}$, namely

$$F_{\mathrm{F},S}(z, w) = \frac{1 + (1-w)\, z}{1 - (2+w)\, z - (1-w)\, z^2}. \tag{16}$$

Equivalently, the coefficients satisfy the recurrence $f_n = (2+w)\, f_{n-1} + (1-w)\, f_{n-2}$ with $f_0 = 1$, $f_1 = 3$. At $w = 0$ this recovers $F_{\mathrm{F},S}(z, 0) = (1+z)/(1-2z-z^2)$, the same denominator as b_n in Section 3.2—which is correct, since both sequences count length-n A-B-S (or C-Y-S) strings with no fluorescent adjacency, regardless of alphabet relabeling. At $w = 1$ the generating function becomes $F_{\mathrm{F},S}(z, 1) = 1/(1-3z)$, so $[z^n]F_{\mathrm{F},S}(z, 1) = 3^n$ as required.

The denominator factors as $(1 - \lambda_+ z)(1 - \lambda_- z)$ where the eigenvalues of $M(w)$ are

$$\lambda_\pm(w) = 1 + \frac{w \pm \sqrt{w^2 + 8}}{2},$$

recovering $\lambda_+(1) = 3$, $\lambda_-(1) = 0$, and $\lambda_\pm(0) = 1 \pm \sqrt{2}$ consistently with the $w = 0, 1$ specializations.

Expected value via $\partial F_{\mathbf{F},S}/\partial w$. Differentiating (16) in w and evaluating at $w = 1$ gives, after simplification,

$$\left.\frac{\partial F_{\mathrm{F},S}}{\partial w}\right|_{w=1} = \frac{2z^2}{(1-3z)^2}.$$

Hence

$$E(n) \cdot 3^n = [z^n]\frac{2z^2}{(1-3z)^2} = 2(n-1)\,3^{n-2}, \qquad n \geq 2,$$

giving $E(n) = 2(n-1)/9$ exactly, with no exponential correction term. This matches the linearity-of-expectation derivation given below.

Verification. For $n = 3$, brute-force enumeration over all $3^3 = 27$ amyloids in $\{C, Y, S\}^3$ gives $\mathcal{A}(3,0) = 17$, $\mathcal{A}(3,1) = 8$, and $\mathcal{A}(3,2) = 2$, with $E(3) = (0{\cdot}17{+}1{\cdot}8{+}2{\cdot}2)/27 = 12/27 = 4/9$. This agrees with both the linearity-of-expectation formula $E(3) = 2(3{-}1)/9 = 4/9$ and the generating-function expansion: $[z^3]F_{\mathrm{F},S}(z,w) = 17 + 8w + 2w^2$ by direct computation of the recurrence $f_n = (2+w)f_{n-1} + (1-w)f_{n-2}$. For $n = 4$, both routes give $\mathcal{A}(4,k) = 41, 28, 10, 2$ for $k = 0, 1, 2, 3$, with $E(4) = (28 + 20 + 6)/81 = 54/81 = 2/3 = 2 \cdot 3/9$ as expected. (See `verify_section5_gf.py`.)

The bivariate generating function above encodes the full probability distribution via $P(n,k) = \mathcal{A}(n,k)/3^n$. However, for computing the expected value, a much simpler route is available. As in Section 4, the total yellow fluorescence of an amyloid equals the number of adjacent $\{C, Y\}$ pairs. Since FRET interactions have no blocking constraint (unlike split-YFP fusion), the linearity of expectation argument from Section 4 applies directly. With each protein independently C, Y, or S with probability 1/3:

$$E(n) = (n-1) \cdot P(\{C,Y\} \text{ pair}) = (n-1) \cdot 2 \cdot \frac{1}{3} \cdot \frac{1}{3} = \frac{2(n-1)}{9}$$

Therefore

$$R_{\mathrm{F},S}(n) := E(n)/n = \frac{2(n-1)}{9n} \longrightarrow \frac{2}{9} \text{ as } n \to \infty$$

The first 400 values of $R_{\mathrm{F},S}(n)$ appear in Figure 4.

6 Variance, null-model fluctuations, and finite-radius FRET

6.1 Finite-n variances

The preceding sections emphasize exact expectations, but the same probability distributions also give exact finite-n fluctuations. Let K_n denote the fluorescence count in the relevant model. If

$$F(z,w) = \sum_{n \geq 0} \sum_{k \geq 0} \mathcal{A}(n,k) w^k z^n$$

is the corresponding bivariate generating function and T_n is the total number of amyloids of length n, then

$$E[K_n] = \frac{[z^n]F_w(z,1)}{T_n}, \qquad E[K_n(K_n-1)] = \frac{[z^n]F_{ww}(z,1)}{T_n},$$

and hence

$$\mathrm{Var}(K_n) = E[K_n(K_n-1)] + E[K_n] - E[K_n]^2.$$

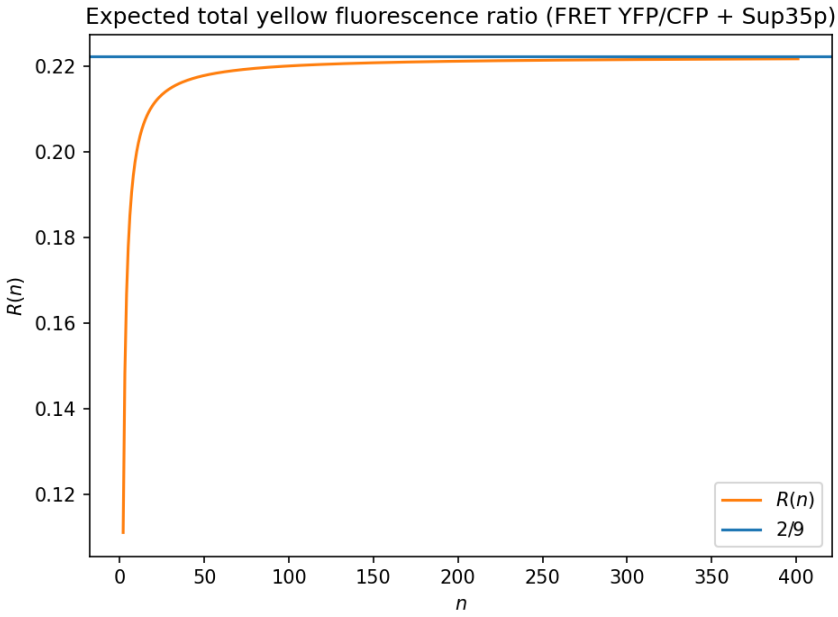

Figure 4: $R_{\mathrm{F},S}(n)$ for FRET YFP/CFP amyloids with Sup35p. The blue line is 2/9 and the orange curve is the exact $R_{\mathrm{F},S}(n) = 2(n-1)/(9n)$. Convergence is algebraic ($O(1/n)$), as in Figure 3.

The variance formulas below are stated for $n \geq 2$; for $n = 1$ one has $K_1 = 0$ and variance 0 in all four models. For reference, the model sizes and bivariate generating functions used in this section are:

Model	T_n	$F(z,w)$
split-YFP	2^n	$\dfrac{1+z}{1-z-2wz^2}$
split-YFP+Sup35p	3^n	$\dfrac{1+z}{1-2z-(1+2w)z^2}$
FRET	2^n	$\dfrac{1+z(1-w)}{1-z(1+w)}$
FRET+Sup35p	3^n	$\dfrac{1+(1-w)z}{1-(2+w)z-(1-w)z^2}$

For the exclusively split-YFP model, the bivariate generating function is

$$F_{\mathrm{SY}}(z,w) = \frac{1+z}{1-z-2wz^2},$$

obtained from the same two-state decomposition as in Section 2. In transfer-matrix form, with p denoting an unfused A/B endpoint and q denoting a fused endpoint, the weighted transition matrix is

$$M_{\mathrm{SY}}(w) = \begin{pmatrix} 1 & w \\ 2 & 0 \end{pmatrix},$$

where the rows and columns are ordered as (p, q) and w marks a fusion. Including the empty amyloid and the initial vector $(2, 0)$ gives

$$F_{\mathrm{SY}}(z, w) = 1 + z(2, 0)\,(I - zM_{\mathrm{SY}}(w))^{-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{1+z}{1-z-2wz^2}.$$

Its second derivative at $w = 1$ is

$$\left.\frac{\partial^2 F_{\mathrm{SY}}}{\partial w^2}\right|_{w=1} = \frac{8z^4}{(1-2z)^3(1+z)^2}.$$

Using this together with the expectation from Section 2 gives, for $n \geq 2$,

$$\mathrm{Var}_{\mathrm{SY}}(K_n) = \frac{6n+2}{81} + \frac{(12n+2)(-1)^n}{81 \cdot 2^n} - \frac{4}{81 \cdot 4^n}.$$

For split-YFP with Sup35p, differentiating (12) twice gives

$$\left.\frac{\partial^2 F}{\partial w^2}\right|_{w=1} = \frac{8z^4}{(1-3z)^3(1+z)^2},$$

and therefore

$$\mathrm{Var}_{\mathrm{SY},S}(K_n) = \frac{56n-27}{576} + \frac{(4n+3)(-1)^n}{48 \cdot 3^n} - \frac{1}{64 \cdot 9^n}, \qquad n \geq 2.$$

For FRET without Sup35p, Section 4 identifies K_n as a Binomial$(n-1, 1/2)$ random variable, so

$$\mathrm{Var}_{\mathrm{F}}(K_n) = \frac{n-1}{4}.$$

For FRET with Sup35p, differentiating (16) twice gives

$$\left.\frac{\partial^2 F_{\mathrm{F},S}}{\partial w^2}\right|_{w=1} = \frac{4z^3(1-z)}{(1-3z)^3},$$

which yields

$$\mathrm{Var}_{\mathrm{F},S}(K_n) = \frac{18n-22}{81}, \qquad n \geq 2.$$

As a small check, at $n = 2$ the split-YFP-with-Sup35p and FRET-with-Sup35p distributions have $P(K_2 = 0) = 7/9$ and $P(K_2 = 1) = 2/9$, giving variance $14/81$, exactly as the formulas above give. In the no-Sup35p split-YFP case, $P(K_2 = 0) = P(K_2 = 1) = 1/2$, while at $n = 3$ one has $P(K_3 = 0) = 2/8$ and $P(K_3 = 1) = 6/8$, giving variances $1/4$ and $3/16$, again matching the formula. The accompanying script `verify_variance.py` brute-force checks all four variance formulas against direct enumeration of $P(n, k)$ for $2 \leq n \leq 7$.

These formulas give exact intrinsic fluctuation scales under the i.i.d. null model; they are not, by themselves, experimental measurement error bars. Since the split-YFP ratios are $2K_n/n$ and the FRET ratios are K_n/n,

$$\mathrm{Var}(R_{\mathrm{SY}}(n)) = \frac{4}{n^2}\,\mathrm{Var}_{\mathrm{SY}}(K_n), \qquad \mathrm{Var}(R_{\mathrm{SY},S}(n)) = \frac{4}{n^2}\,\mathrm{Var}_{\mathrm{SY},S}(K_n)$$

$$\mathrm{Var}(R_{\mathrm{F}}(n)) = \frac{1}{n^2}\,\mathrm{Var}_{\mathrm{F}}(K_n), \qquad \mathrm{Var}(R_{\mathrm{F},S}(n)) = \frac{1}{n^2}\,\mathrm{Var}_{\mathrm{F},S}(K_n)$$

Thus, for large n, the standard deviations of the ratio statistics scale as $O(n^{-1/2})$. For example,

$$\mathrm{Var}(R_{\mathrm{SY}}(n)) \sim \frac{8}{27n}, \quad \mathrm{Var}(R_{\mathrm{SY},S}(n)) \sim \frac{7}{18n}, \quad \mathrm{Var}(R_{\mathrm{F}}(n)) \sim \frac{1}{4n}, \quad \mathrm{Var}(R_{\mathrm{F},S}(n)) \sim \frac{2}{9n}.$$

This is the usual central-limit scale for additive statistics of finite-state chains and transfer matrices. More formal Gaussian limit laws could be obtained from standard transfer-matrix or quasi-power arguments, but the exact finite-n distributions and variances above are the quantities used here. For short amyloids, the exact distributions $P(n,k)$ should be used directly. For experimental ensembles of many fibrils, the relevant uncertainty also depends on the sampling design, the number of fibrils measured, and instrumental noise.

6.2 Finite-radius FRET expectation

The nearest-neighbor FRET assumption can also be relaxed at the level of expected value. Suppose a YFP can receive FRET from CFPs within distance r along the amyloid, with additive unit contribution from each CFP–YFP pair in this window. Let p_C and p_Y be the probabilities of CFP and YFP, and set $m = \min(r, n-1)$. There are $n-d$ unordered pairs at distance d, each contributing in expectation $2p_Cp_Y$. Linearity of expectation requires no independence among these pair indicators, so the expectation is exact even though the indicators are correlated. Hence

$$E_r(n) = 2p_Cp_Y \sum_{d=1}^{m} (n-d) = 2p_Cp_Y \left(mn - \frac{m(m+1)}{2} \right).$$

For fixed r and $n \to \infty$, this gives

$$\frac{E_r(n)}{n} \longrightarrow 2rp_Cp_Y.$$

In the equal-concentration FRET+Sup35p case ($p_C = p_Y = 1/3$), the nearest-neighbor model $r = 1$ recovers $E_1(n) = 2(n-1)/9$, while $r = 2$ gives

$$E_2(n) = \frac{2}{9}\big((n-1) + (n-2)\big) = \frac{4n-6}{9}, \qquad n \geq 3,$$

and $E_2(n)/n \to 4/9$. A crude hard-cutoff surrogate motivated by $R_0/0.47\,\mathrm{nm} \approx 10$ would scale this unit-window baseline linearly in r, but that number is a property of the idealized indicator sum, not a FRET-efficiency-weighted prediction. The point is only that the nearest-neighbor value $2/9$ is an $r = 1$ mathematical baseline rather than a quantitative experimental prediction.

The unit-weight window is itself an idealization: a more physical model could weight a CFP–YFP pair at distance d by a distance-dependent FRET efficiency η_d rather than by a hard cutoff (for example, an efficiency kernel depending on distance and orientation). The expected-value computation would still go through, giving

$$\frac{E(n)}{n} \longrightarrow 2p_Cp_Y \sum_{d \geq 1} \eta_d$$

whenever the efficiency weights are summable, but the closed form would depend on the chosen geometry. Computing the full distribution for $r > 1$ is substantially harder: adjacent indicators are correlated over a larger window, and a transfer matrix would need to track the last r protein identities rather than only the rightmost state. For $n = 2$, the value $m = \min(r, n-1)$ is still 1 even when $r = 2$, so the radius-two case collapses to the nearest-neighbor formula until $n \geq 3$.

7 Generalization to unequal concentrations

The results of Sections 2–5 all depend on the equal-concentration assumption. The Markov chain arguments generalize readily to unequal concentrations, which we outline here. This section records limiting per-step fluorescence ratios rather than full finite-n distributions. Such distributions could be obtained by the same finite-state transfer-matrix method, with probabilities replacing the equal-count weights, but the resulting formulas are less compact and are left as a natural extension for experimental calibration.

7.1 Split-YFP with unequal concentrations

Suppose protein A appears with probability α and protein B with probability $1-\alpha$. A two-state chain on {unfused, fused} is insufficient here, because the probability that the next protein triggers a fusion conditional on the current rightmost being unfused depends on whether the rightmost is an A or a B: from rightmost-A, fusion requires next = B (probability $1-\alpha$), while from rightmost-B it requires next = A (probability α). For $\alpha \neq 1/2$ these are not equal, so we cannot lump A-unfused and B-unfused into a single state without loss. We instead use the three-state chain $\{p_A, p_B, q\}$ from Section 7.2, specialized to the no-Sup35p case $\beta = 1-\alpha$ (so $1-\alpha-\beta = 0$). The transitions are:

- From p_A: next is A (prob α, stay p_A); next is B (prob $1-\alpha$, fuse, go q).
- From p_B: next is B (prob $1-\alpha$, stay p_B); next is A (prob α, fuse, go q).
- From q: next is A (prob α, go p_A); next is B (prob $1-\alpha$, go p_B).

The balance equations $\pi_A = \alpha\pi_A + \alpha\pi_q$ and $\pi_B = (1-\alpha)\pi_B + (1-\alpha)\pi_q$, together with $\pi_A + \pi_B + \pi_q = 1$, give

$$\pi_A = \frac{\alpha^2}{1-\alpha(1-\alpha)}, \qquad \pi_B = \frac{(1-\alpha)^2}{1-\alpha(1-\alpha)}, \qquad \pi_q = \frac{\alpha(1-\alpha)}{1-\alpha(1-\alpha)}.$$

The expected fluorescence per step is $e = (1-\alpha)\pi_A + \alpha\pi_B = \alpha(1-\alpha)/(1-\alpha(1-\alpha))$, so

$$\lim_{n\to\infty} R_{\mathrm{SY}}(n) = 2e = \frac{2\alpha(1-\alpha)}{1-\alpha(1-\alpha)}.$$

Equivalently, this is the specialization $\beta = 1-\alpha$ of the Section 7.2 formula $2\alpha\beta(2-\alpha-\beta)/(1-\alpha\beta)$. At $\alpha = 1/2$, $R_{\mathrm{SY}} \to (1/2)/(3/4) = 2/3$, recovering Section 2. The maximum occurs at equal concentrations: differentiating gives

$$\frac{d}{d\alpha}\left(\frac{2\alpha(1-\alpha)}{1-\alpha(1-\alpha)}\right) = \frac{2(1-2\alpha)}{(1-\alpha(1-\alpha))^2},$$

so the only interior critical point is $\alpha = 1/2$, and the endpoints have zero fluorescence. Numerically, at $\alpha = 1/4$ the limit is $2 \cdot \frac{3}{16}/\frac{13}{16} = 6/13 \approx 0.462$, in agreement with Monte Carlo simulation (the default script run with $n = 5000$ and 400 trials gives approximately 0.462; see `split_yfp_unequal_simulation.py`). A naive two-state chain would wrongly lump p_A and p_B into a single unfused state and assign fusion probability $d = 2\alpha(1-\alpha)$ from that lumped state, giving the ratio $2d/(1+d) = 4\alpha(1-\alpha)/(1+2\alpha(1-\alpha))$; at $\alpha = 1/4$ this is $6/11 \approx 0.545$. The two-state lumping is harmless at $\alpha = 1/2$, where A/B symmetry gives equal transition probabilities from p_A and p_B and hence a genuinely lumpable chain, but it fails the simulation check at every $\alpha \neq 1/2$.

The convergence rate of $R_{\mathrm{SY}}(n)$ to its limit depends on α. Near $\alpha = 0$ or $\alpha = 1$, the chain changes state rarely and finite-n values can differ noticeably from the limiting ratio. For experimental fibril lengths, the finite-n value obtained by iterating the three-state chain or by simulation should be used rather than the stationary limit alone.

7.2 Split-YFP with Sup35p at unequal concentrations

Now suppose proteins A, B, S appear with probabilities α, β, and $1-\alpha-\beta$ respectively. We define three Markov states: p_A (rightmost is A, unfused), p_B (rightmost is B, unfused), and q (rightmost is S, or fused). The transitions are:

- From p_A: next is A (prob α, stay p_A), B (prob β, fuse, go q), or S (prob $1-\alpha-\beta$, go q).
- From p_B: next is B (prob β, stay p_B), A (prob α, fuse, go q), or S (prob $1-\alpha-\beta$, go q).
- From q: next is A (prob α, go p_A), B (prob β, go p_B), or S (prob $1-\alpha-\beta$, stay q).

Solving the balance equations $\pi_A = \alpha\pi_A + \alpha\pi_q$, $\pi_B = \beta\pi_B + \beta\pi_q$, and $\pi_A + \pi_B + \pi_q = 1$ gives $\pi_A = \alpha(1-\beta)/(1-\alpha\beta)$, $\pi_B = \beta(1-\alpha)/(1-\alpha\beta)$, and $\pi_q = (1-\alpha)(1-\beta)/(1-\alpha\beta)$. The expected fluorescence per step is $e = \beta\pi_A + \alpha\pi_B = \alpha\beta(2-\alpha-\beta)/(1-\alpha\beta)$, giving

$$\lim_{n\to\infty} R_{\mathrm{SY},S}(n) = 2e = \frac{2\alpha\beta(2-\alpha-\beta)}{1-\alpha\beta}$$

which reduces to $1/3$ when $\alpha = \beta = 1/3$ (and to $2/3$ when $\alpha = \beta = 1/2$, recovering the no-Sup35p case). The same simulation script also checks this unequal-Sup35p limit for representative (α, β) pairs.

7.3 FRET with unequal concentrations

In the two-species FRET setting, $P(C) = p_C$ and $P(Y) = p_Y = 1 - p_C$, so linearity of expectation gives $R_{\mathrm{F}}(n) \to 2p_Cp_Y$ directly. In the three-species FRET+Sup35p setting, $P(C) = p_C$, $P(Y) = p_Y$, and $P(S) = p_S$ with $p_C + p_Y + p_S = 1$; the same adjacent-pair calculation gives $R_{\mathrm{F},S}(n) \to 2p_Cp_Y$. These generalizations would allow the model to be calibrated to experimental concentration ratios.

8 Conclusion

We have derived exact probability distributions, closed-form expected values, and finite-n variances for fluorescence in four combinatorial models of amyloid fibrils. The exact expectation formulas below are for $n \geq 1$, with $E(1) = 0$ in all four models; the $n = 0$ terms used in generating functions are formal conventions only. The principal expectation results are summarized below in each model's native units:

System	Native ratio	Exact $E(n)$	Exact ratio	Limit
Split-YFP (Sec. 2)	$R_{\mathrm{SY}} = \dfrac{2E}{n}$	$\dfrac{3n-2}{9} + \dfrac{2(-1)^n}{9 \cdot 2^n}$	$\dfrac{2(3n-2)}{9n} + \dfrac{2(-1)^n}{9n \cdot 2^{n-1}}$	2/3
Split-YFP + Sup35p (Sec. 3)	$R_{\mathrm{SY},S} = \dfrac{2E}{n}$	$\dfrac{4n-3}{24} - \dfrac{(-1)^{n-1}}{24 \cdot 3^{n-1}}$	$\dfrac{4n-3}{12n} + \dfrac{(-1)^n}{12n \cdot 3^{n-1}}$	1/3
FRET YFP/CFP (Sec. 4)	$R_{\mathrm{F}} = \dfrac{E}{n}$	$\dfrac{n-1}{2}$	$\dfrac{n-1}{2n}$	1/2
FRET + Sup35p (Sec. 5)	$R_{\mathrm{F},S} = \dfrac{E}{n}$	$\dfrac{2(n-1)}{9}$	$\dfrac{2(n-1)}{9n}$	2/9

Here R_{SY} and $R_{\mathrm{SY},S}$ are fractions of proteins participating in split-YFP fluorescence (two proteins per fused pair), while R_{F} and $R_{\mathrm{F},S}$ are total FRET fluorescence units per protein. The table places the four exact baselines side by side for reference, but the split-YFP and FRET ratios measure different physical quantities and are not directly comparable.

A key structural distinction separates the two pairs of systems. In the split-YFP settings (Sections 2 and 3), the adjacency blocking constraint—a fused pair cannot participate in further fusion—introduces nontrivial combinatorial structure. This necessitates generating function techniques to derive the full distributions $P(n, k)$ and Markov chain arguments to compute the expected values in closed form. The agreement between these two independent approaches provides mutual verification of both the sum identities and the Markov chain analysis.

In the FRET settings (Sections 4 and 5), the absence of a blocking constraint means that the total fluorescence is a sum of adjacent-pair contributions: one interaction does not consume a protein and prevent another interaction. Thus a direct linearity-of-expectation argument gives the exact expectation. In the two-letter FRET model these adjacent-pair indicators are independent by the bijection in Section 4, while in the Sup35p model they are correlated, as reflected in the variance formula.

The limiting ratios admit clean interpretations under the i.i.d. null model. For split-YFP amyloids under random assembly, $R_{\mathrm{SY}}(n) \to 2/3$ means that in a long amyloid composed of equal concentrations of the two halves, approximately two-thirds of the proteins participate in a fluorescent fusion—a high baseline fluorescence yield arising from the alternating tendency of random binary sequences. Introducing Sup35p at equal concentration gives $R_{\mathrm{SY},S}(n) \to 1/3$: the inert Sup35p proteins act as spacers that interrupt potential fusion pairs. For FRET, the ratio $R_{\mathrm{F}}(n) \to 1/2$ reflects the null-model probability that a random adjacent pair is a $\{C, Y\}$ pair, while $R_{\mathrm{F},S}(n) \to 2/9$ reflects the reduced probability $2 \cdot (1/3)^2$ of such a pair when three protein types compete for each position. The variance formulas in Section 6 give finite-n null-model fluctuation scales, so experimentally measured fluorescence ratios can be compared against both the expected value and the size of random fluctuations under the i.i.d. model.

Section 6 also records the expected-value extension of the FRET model to an interaction radius r. This partially addresses the nearest-neighbor simplification while preserving a

closed-form baseline: the expectation remains a direct sum over all CFP–YFP pairs within distance r. A hard cutoff with r on the order of ten amyloid spacings would scale the unit-window expectation far above the nearest-neighbor value, underscoring that the $r = 1$ model is a mathematically transparent baseline rather than a quantitative FRET prediction. However, the full distribution for $r > 1$ would require generating functions or transfer matrices for strings with longer-range dependencies, a substantially richer combinatorial problem.

Similarly, Section 7 gives limiting ratios for unequal concentrations but not the corresponding finite-n distributions or variances. Those distributions are accessible in principle through larger or weighted transfer matrices, and would be the next step for fitting the null model to measured concentration ratios.

Our analysis assumes that proteins are independently distributed in the amyloid sequence. This assumption neglects known cooperative and templated growth effects in amyloid assembly [1] and should be interpreted as a combinatorial null model: a baseline against which experimental deviations—reflecting assembly biases, kinetic trapping, or cooperative interactions—could be quantified. The combinatorial framework developed here, particularly the generating function approach to constrained fluorescence patterns via regular expressions, could in principle be adapted to non-uniform protein distributions if experimental data on assembly biases become available.

All simulation and verification scripts referenced in this paper are available at

`https://github.com/danielchen0/amyloids`.

These include the split-YFP, Sup35p, FRET, unequal-concentration, variance, and generating function verification scripts, together with the associated plotting scripts.

References

[1] Fabrizio Chiti and Christopher M. Dobson. Protein misfolding, functional amyloid, and human disease. *Annual Review of Biochemistry*, 75:333–366, 2006. `https://doi.org/10.1146/annurev.biochem.75.101304.123901`.

[2] Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2009.

[3] Theodor Förster. Zwischenmolekulare Energiewanderung und Fluoreszenz. *Annalen der Physik*, 437(1–2):55–75, 1948. `https://doi.org/10.1002/andp.19484370105`.

[4] Ronald L. Graham, Donald E. Knuth, and Oren Patashnik. *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley, 2nd edition, 1994.

[5] Chang-Deng Hu, Yurii Chinenov, and Tom K. Kerppola. Visualization of interactions among bZIP and Rel family proteins in living cells using bimolecular fluorescence complementation. *Molecular Cell*, 9(4):789–798, 2002. `https://doi.org/10.1016/S1097-2765(02)00496-3`.

[6] Tom K. Kerppola. Design and implementation of bimolecular fluorescence complementation (BiFC) assays for the visualization of protein interactions in living cells. *Nature Protocols*, 1(3):1278–1286, 2006. `https://doi.org/10.1038/nprot.2006.201`.

[7] Joseph R. Lakowicz. *Principles of Fluorescence Spectroscopy*. Springer, 3rd edition, 2006. https://doi.org/10.1007/978-0-387-46312-4.

[8] James R. Norris. *Markov Chains*. Cambridge University Press, 1997.

[9] OEIS Foundation Inc. A000129. https://oeis.org/A000129. Accessed 2026-05-14.

[10] OEIS Foundation Inc. A001333. https://oeis.org/A001333. Accessed 2026-05-14.

[11] OEIS Foundation Inc. A095977. https://oeis.org/A095977. Accessed 2026-05-14.

[12] OEIS Foundation Inc. A127976. https://oeis.org/A127976. Accessed 2026-05-14.

[13] M. M. Patino, J. J. Liu, J. R. Glover, and S. Lindquist. Support for the prion hypothesis for inheritance of a phenotypic trait in yeast. *Science*, 273(5275):622–626, 1996. https://doi.org/10.1126/science.273.5275.622.

[14] George H. Patterson, David W. Piston, and B. George Barisas. Förster distances between green fluorescent protein pairs. *Analytical Biochemistry*, 284(2):438–440, 2000. https://doi.org/10.1006/abio.2000.4708.

[15] Richard P. Stanley. *Enumerative Combinatorics*, volume 1. Cambridge University Press, 2nd edition, 2012.

[16] Margaret Sunde, Louise C. Serpell, Mark Bartlam, Peter E. Fraser, Mark B. Pepys, and Colin C. F. Blake. Common core structure of amyloid fibrils by synchrotron x-ray diffraction. *Journal of Molecular Biology*, 273(3):729–739, 1997. https://doi.org/10.1006/jmbi.1997.1348.

[17] Herbert S. Wilf. *generatingfunctionology*. A K Peters, 3rd edition, 2006.