Coagulation-fragmentation with a finite number of particles: models, stochastic analysis and applications to telomere clustering and viral capsid assembly

N. Hoze [1] and D. Holcman [2]

Abstract

Coagulation-fragmentation processes with a finite number of particles is a recent class of mathematical questions that serves modeling some cell biology dynamics. The analysis of the models offers new challenging questions in probability and analysis: the model is the clustering of particles after binding, the formation of local subclusters of arbitrary sizes and the dissociation into subclusters. We review here modeling and analytical approaches to compute the size and number of clusters with a finite size. Applications are clustering of chromosome ends (telomeres) in yeast nucleus and the formation of viral capsid assembly from molecular components. The methods to compute the probability distribution functions of clusters and to estimate the statistical properties of clustering are based on combinatorics and hybrid Gillespie-spatial simulations. Finally, we review models of capsid formation, the mean-field approximation and jump processes used to compute first passage times to a finite size cluster. These models become even more relevant for extracting parameters from live cell imaging data.

1 Introduction

Clustering processes are generic in statistical physics and biology. For example in astrophysics, masses can form aggregate under the gravitation force, while in biochemistry, molecules interact to form colloids that aggregate in solution [6]. In cell biology, aggregation underlies beta-amyloid structure formation involved in Alzheimer disease or chromosomal organization in the cell nucleus. However a new class of mathematical problems appears with the need to analyze clustering with a finite number of random particles such as the organization of the chromosome ends [17] or viral capsid assembly in cells. These processes are modeled as coagulation-fragmentation.

Irreversible aggregation of many particles in clusters was already described by Von Smoluchowski in 1916 [36] to model an infinite number

[1]Institut für Integrative Biologie, ETH, Universitätstrasse 16, 8092 Zürich, Switzerland.

[2]Applied Mathematics and Computational Biology, Ecole Normale Supérieure, 46 rue d'Ulm 75005 Paris, France and Mathematical Institute, University of Oxford, Oxford OX2 6GG, United Kingdom. UK.

of interacting molecules. When a cluster can lose or gain only one particle at a time, the Smoluchowski equations become the Becker-Döring model which consists in an ensemble of coagulation-fragmentation equations [5, 6, 28, 40]. Nowadays, determinist, stochastic, asymptotic and numerical methods are developed to study steady-state and transient properties of clustering based on molecular components [2, 10, 39, 33, 7]. Another class of problem concerns the clustering with an infinite number of particles (Marcus-Lushnikov process) [30, 31, 25], but much less is known about coagulation-fragmentation with a finite number of particles [14]. When the cluster size cannot exceed a given threshold, new difficulties arises in the analysis of the coagulation-fragmentation equations [17, 41]. These models are relevant in molecular genetics for characterizing the organization of the chromosome ends [17] or to model viral capsid assembly in cell biology [42, 21, 22].

We review here several models, asymptotic and combinatorial results as well as a generalization of the Gillespie's algorithm to study aggregation in spatially inhomogeneous environment. In the first section, we describe the Smoluchowski equations for coagulation-fragmentation. In the second, we present a general analysis and result about clustering with a finite number of particle. Section three is dedicated to Gillespie's algorithm in spatially inhomogeneous environment, applied to telomere organization in yeast. In sections four and five, we present asymptotic methods for capsid viral assembly and the analysis of single particle trajectories.

2 Primer in Smoluchowski equations for coagulation-fragmentation

2.1 Coagulation-fragmentation with an infinite number of particles

This section summarizes the Smoluchowski equations for coagulation-fragmentation that consist of an infinite system of differential equations for the number $n(j,t)$ of clusters of size j at time t in a population of infinite size [36]. The coagulation process is characterized by the rate $C(i,j)$ by which two clusters of size i and j coalesce to form a cluster of size $i+j$, while fragmentation with rate $F(i,j)$ describes that a cluster of size $i+j$ dissociates into a cluster of size i and a cluster of size j. The conservation of mass equation is given

by

$$\begin{aligned}\frac{dn(j,t)}{dt} &= \frac{1}{2}\sum_{k=0}^{j-1} C(k,j-k)n(k,t)n(j-k,t) - n(j,t)\sum_{k=1}^{\infty} C(j,k)n(k,t) \\ &- n(j,t)\sum_{k=1}^{j-1} F(k,j-k) + \sum_{k=1}^{\infty} F(j,k)n(k+j,t), \qquad (1)\end{aligned}$$

where the index j can take values between 1 and ∞ and the first line in the left-hand side corresponds to the coagulation and the second accounts for the fragmentation. This system of equations is a mean-field deterministic model of the coagulation-fragmentation process that do not describe intrinsic cluster interactions.

Coagulation-fragmentation processes (CFP)satisfies the balance condition [13], for which there exists a function $a(i) = a_i$ such that $\forall i, j \in \mathbb{N}$

$$\frac{C(i,j)}{F(i,j)} = \frac{a(i+j)}{a(i)a(j)}. \qquad (2)$$

When the total number of clusters is fixed, the probability distribution function of the number of clusters can be computed, as well as the probability distribution that the number of cluster of size i is m_i so that the distribution of sizes of clusters is $(m_1, \ldots, m_n)$. When there are exactly N particles and the total number of clusters is fixed to K, the following identity for number conservation is satisfied [27]

$$\sum_{i=1}^{N} m_i = K. \qquad (3)$$

When the total number of clusters is K, the conditional probability distribution function is given by

$$p'(m_1, \ldots, m_N|K) = \frac{1}{C_{N,K}} \frac{a(1)^{m_1} \ldots a(N)^{m_N}}{m_1!...m_N!},$$

where the normalization constant $C_{N,K}$ will be described below (see formula 53). This formulas are used to compute the statistical moments for the cluster distributions.

2.2 Continuous-time Markov chain equations for a finite number of particles

The steady-state distribution for a CFP stochastic model with a finite number of N particles is described by a continuous-time Markov chain equations

in the cluster configuration space. We start with N particles distributed in clusters of size $(n_1, \ldots, n_K)$ that can undergo coagulation or fragmentation events under the constraint that

$$\sum_{k=1}^{K} n_k = N. \tag{4}$$

To study the distribution of particles in clusters, we use the decomposition of the integer N in a sum of positive integers (integer partition) [3]. The partitions of the integer N are described in dimension N by the ensemble

$$P_N = \{(n_1, \ldots, n_N) \in \mathbb{N}^N; \sum_{i=1}^{N} n_i = N \text{ and } n_1 \geq \cdots \geq n_N \geq 0\}. \tag{5}$$

The probability $P(n_1, \ldots, n_N, t)$ of the configuration $(n_1, \ldots, n_N)$ at time t, satisfies an ensemble of close equations obtained by considering all possible coagulations or fragmentations between time t and $t + \Delta t$:

- Two clusters of size n_i and n_j coagulates with a probability $C(n_i, n_j)\Delta t$ to form a cluster of size $n_i + n_j$.
- A cluster of size n_i dissociates into two clusters of size k and $n_i - k$ with a probability $F(k, n_i - k)\Delta t$.
- Nothing happens with the probability $1 - \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} C(n_i, n_j)\Delta t - \sum_{i=1}^{N} \sum_{k=1}^{n_i - 1} F(k, n_i - k)\Delta t$.

Thus, the probability $P(n_1, \ldots, n_N, t)$ satisfies

$$\begin{aligned}
\frac{d}{dt} P(n_1, \ldots, n_N, t) &= -\left(\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} C(n_i, n_j) + \sum_{i=1}^{N} \sum_{k=1}^{n_i - 1} F(k, n_i - k) \right) P(n_1, \ldots, n_N, t) \\
&+ \sum_{k=1}^{N} \sum_{\substack{n_i' > 0, n_j' > 0 \\ n_i' + n_j' = n_k}} C(n_i', n_j') P(n_1, \ldots, n_i', \ldots, n_j', \ldots, n_N, t) \\
&+ \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} F(n_i, n_j) P(n_1, \ldots, n_i + n_j, \ldots, n_N, t).
\end{aligned} \tag{6}$$

Moreover, $C(n_i, n_j) = 0$ if either n_i or n_j is equal to 0. The partitions of the integer N are described by the set

$$P_N = \{(n_1, \ldots, n_N) \in \mathbb{N}^N; \sum_{i=1}^{N} n_i = N \text{ and } n_1 \geq \cdots \geq n_N \geq 0\} \tag{7}$$

and the ensemble of decompositions

$$P'_N = \{(m_1, \ldots, m_N) \in \mathbb{N}^N; \sum_{i=1}^{N} i m_i = N \text{ and } n_1, \ldots, n_N \geq 0\}. \tag{8}$$

In the ensemble P'_N, m_i is the number of occurrence of integer i in the partition of the integer N. The two ensembles P_N and P'_N corresponds to different representations of the clusters distributions.

For example $N = 9$ particles are distributed in two clusters of one particle, two clusters of two, and one cluster of three and the distribution is $(3, 2, 2, 1, 1, 0, 0, 0, 0) \in P_9$, and $(2, 2, 1, 0, 0, 0, 0, 0, 0) \in P'_9$.

When the coefficient C and F satisfies relation 2, there exists an invariant measure [13] for the steady-state probability of a given configuration $(m_1, \ldots, m_N) \in P'_N$, given by

$$P'(m_1, \ldots, m_N) = \frac{1}{C_N} \frac{a_1^{m_1} \ldots a_N^{m_N}}{m_1! \ldots m_N!}, \tag{9}$$

where C_N is a normalization constant. Computing the normalization constant explicitly is difficult [38]. In the next subsection, we estimate the probability of occurrence of a certain cluster configuration $(m_1, \ldots, m_N)$.

2.3 Description of the cluster partitions with a finite number of particles

To determine the cluster distribution at equilibrium, we compute here the probability of a configuration when the number of clusters K is fixed. We also find the probability of having K clusters. The number of distributions of N particles into K clusters is the cardinal of the ensemble

$$P_{N,K} = \{(n_1, \ldots, n_K) \in (\mathbb{N})^N; \sum_{i=1}^{K} n_i = N \text{ and } n_1 \geq \cdots \geq n_N \geq 0\}, \tag{10}$$

which is also the ensemble of the partitions of the integer N as a sum of K integers. This ensemble is in bijection with

$$P'_{N,K} = \{(m_1, \ldots, m_N) \in \mathbb{N}^N; \sum_{i=1}^{N} i m_i = N \text{ and } \sum_{i=1}^{N} m_i = K\}, \tag{11}$$

where the application $P_{N,K} \to P'_{N,K}$ defined by

$$(n_1, \ldots, n_K) \mapsto (m_1, \ldots, m_N) = (\sum_{i=1}^{K} 1_{\{n_i=1\}}, \ldots, \sum_{i=1}^{K} 1_{\{n_i=N\}}) \tag{12}$$

maps the partition $(n_1, \ldots, n_N)$ where N is written as a sum of K positive integers to the number of occurrence of each integer into the image partition. The partitions of N are written as

$$P_N = \bigcup_K P_{N,K} \text{ and } P'_N = \bigcup_K P'_{N,K}. \tag{13}$$

In section 3.1, 3.2 and 3.3, we derive explicitly expressions for the probabilities of configurations in $P'_{N,K}$.

2.4 Statistical moments for the cluster configurations when the number of clusters is fixed

The probability of a configuration $(m_1, \ldots, m_N)$, when the total number of clusters is equal to K, is

$$p'(m_1, \ldots, m_N|K) = \frac{\frac{a_1^{m_1} \ldots a_N^{m_N}}{m_1! \ldots m_N!}}{C_{N,K}}. \tag{14}$$

where

$$C_{N,K} = \sum_{(m_i) \in P'_{N,K}} \frac{a_1^{m_1} \ldots a_N^{m_N}}{m_1! \ldots m_N!}. \tag{15}$$

The normalization factor of eq. (14) is computed using the partial sums

$$S_N(x) = \sum_{i=1}^{N} a_i x^i. \tag{16}$$

The functions S^K and S_N^K have the same N^{th} order coefficient and this coefficient determines $C_{N,K}$. We recall [24] the

Theorem 2.1 *When the number of clusters is equal to K for a total of N particles, the mean number of clusters of size i is*

$$\langle M_i \rangle_{N,K} = a_i \frac{C_{N-i,K-1}}{C_{N,K}}, \tag{17}$$

where a_i and $C_{N,K}$ are defined in (2) *and* (15) *respectively.*

Furthermore, $\langle M_i \rangle_{N,K} = 0$ if $i > N - K + 1$. Interestingly,

Theorem 2.2 *The second moment of the number of clusters of size i is*

$$\begin{aligned} \langle M_i^2 \rangle_{N,K} &= \frac{1}{C_{N,K}} \sum_{P'_{N,K}} m_i^2 \frac{a_1^{m_1} ... a_N^{m_N}}{m_1! ... m_N!} \\ &= a_i^2 \frac{C_{N-2i,K-2}}{C_{N,K}} + a_i \frac{C_{N-i,K-1}}{C_{N,K}}, \end{aligned} \tag{18}$$

and the covariance is

$$\langle M_{i,j}^2 \rangle_{N,K} - \langle M_j^2 \rangle_{N,K} \langle M_j \rangle_{N,K} = a_i a_j \left(\frac{C_{N-i-j,K-2}}{C_{N,K}} - \frac{C_{N-i,K-1} C_{N-j,K-1}}{C_{N,K}^2} \right). \tag{19}$$

The proofs can be found in [24].

2.5 Distribution of the number of clusters

In the previous section, we introduce the probability distribution of a cluster configuration and the statistical moments for a fix number of clusters. In this section, we describe the statistics of the entire cluster configurations using the probability distribution of the *number of clusters*. Our goal is to study the time dependent probability density function

$$P_K(t) = P\{K \text{ clusters at time } t\}, \tag{20}$$

which is associated to a birth-and-death process: the probability of having K clusters at time $t + \Delta t$ is the sum of the probability of starting at time t with $K - 1$ clusters and one of them dissociates into two smaller ones plus the probability of starting with $K + 1$ clusters and two of them associate plus the probability of starting with K and nothing happens (Fig. 1).

The first probability is the product of P_{K-1} by the transition rate $s_{K-1}\Delta t$ to go from state with $K-1$ clusters to K, while the second is the transition from $K+1$ to K, which is the product of P_{K+1} by the transition rate $f_{K+1}\Delta t$ of going from $K + 1$ clusters to K. The master equations are given by

$$\begin{cases} \dot{P}_1(t) &= -s_1 P_1(t) + f_2 P_2(t) \\ \dot{P}_K(t) &= -(f_K + s_K) P_K(t) + f_{K+1} P_{K+1}(t) + s_{K-1} P_{K-1}(t) \\ \dot{P}_N(t) &= -f_N P_N(t) + s_{N-1} P_{N-1}(t). \end{cases} \tag{21}$$

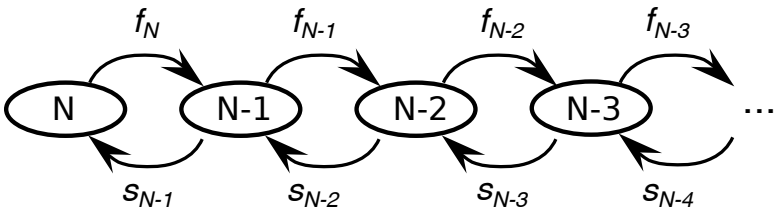

Figure 1: **Markov chain representation for the number of clusters.** s_K (respectively f_K) is the separation (respectively formation) rate of a cluster when there are K clusters.

The steady probability is defined by

$$\Pi_K = \lim_{t\to\infty} P_K(t) \tag{22}$$

where there are K clusters at steady state. The steady state probabilities of the number of clusters are solutions of the system

$$\begin{cases} 0 &= -s_1\Pi_1 + f_2\Pi_2 \\ 0 &= -(f_K + s_K)\Pi_K + f_{K+1}\Pi_{K+1} + s_{K-1}\Pi_{K-1} \\ 0 &= -f_N\Pi_N + s_{N-1}\Pi_{N-1}, \end{cases} \tag{23}$$

with the normalization condition

$$\sum_{K=1}^{N} \Pi_K = 1. \tag{24}$$

The probabilities Π_K are given by the ratio

$$\frac{\Pi_K}{\Pi_{K-1}} = \frac{s_{K-1}}{f_K} \text{ for } K \geq 2 \tag{25}$$

and the coefficients s_K and f_K are the mean-field separation and formation rates respectively. Whereas the cluster configurations when the number of clusters is fixed depend only on the kernel a_i, the statistics of the number of clusters depends on the cluster fragmentation and coagulation rates F and C.

In the following, we will focus on the coagulation condition $C(i,j) = 1$ and the fragmentation $F(i,j) = \frac{a_i a_j}{a_{i+j}}$ to state the

Theorem 2.3 *When $C(i,j) = 1$ and $F(i,j) = \frac{a_i a_j}{a_{i+j}}$, the separation rate when there are K clusters is given by*

$$s_K = \frac{\sum_{i=1}^{N} \sum_{j=1}^{i-1} a_j a_{i-j} C_{N-i,K-1}}{C_{N,K}} \tag{26}$$

and the formation rate when there are K clusters is

$$f_K = \frac{K(K-1)}{2}. \tag{27}$$

Using these results, we can now describe the statistics of the entire cluster configurations. Using Bayes rule, the probability of a configuration $(m_1, \ldots, m_N)$, that contains K clusters is the product of the conditional probability $p'(m_1, \ldots, m_N|K)$ by the probability of having K clusters

$$p'(m_1, \ldots, m_N, K) = p'(m_1, \ldots, m_N|K)\Pi_K. \tag{28}$$

The mean number of clusters of size i is thus

$$\langle M_i \rangle_N = \sum_{K=1}^{N} \Pi_K \langle M_i \rangle_{N,K}. \tag{29}$$

2.6 The probability to find two particles in the same cluster

When the mean number of clusters has reached its equilibrium, particles can still be exchanged between clusters. This exchange is characterized by the probability to find two particles in the same cluster.
When the distribution of the clusters is $(n_1, \ldots, n_K)$, the probability $P_2(n_1, \ldots, n_K)$ to find two given particles in the same cluster is obtained by using the probability to choose the first particle in the cluster n_i, which is equal to the number of particles in the cluster divided by the total number of particles $\frac{n_i}{N}$. The probability to have the second particle in the same cluster is $\frac{n_i - 1}{N-1}$. Summing over all possibilities, we get

$$P_2(n_1, \ldots, n_K) = \sum_{i=1}^{K} \frac{n_i}{N} \frac{n_i - 1}{N-1} = \frac{1}{N(N-1)} (\sum_{i=1}^{K} n_i^2 - N). \tag{30}$$

We note that

$$\sum_{j=1}^{K} n_j^2 = \sum_{i=1}^{N} i^2 m_i, \tag{31}$$

thus we get

$$\begin{aligned}\sum_{(n_1,\ldots,n_K)\in P_{N,K}} p(n_1,\ldots,n_K)\sum_{j=1}^{K} n_j^2 &= \sum_{(m_i)\in P'_{N,K}} p(m_i)\sum_{j=1}^{N} j^2 m_j \\ &= \sum_{j=1}^{N} j^2 \sum_{(m_i)\in P'_{N,K}} m_j p(m_i) \quad (32)\\ &= \sum_{j=1}^{N} j^2 \langle M_j\rangle_{N,K},\end{aligned}$$

where $\langle M_j\rangle_{N,K}$ is the mean number of clusters of size j, when there are N particles distributed in K clusters eq. (17). Taking into account all possible distributions of clusters, we obtain that the probability $\langle P_2\rangle$ to find two particles in the same cluster is

$$\langle P_2\rangle = \sum_{K=1}^{N} \sum_{(n_1,\ldots,n_K)\in P_{N,K}} P_2(n_1,\ldots,n_K) p(n_i)\Pi_K, \quad (33)$$

which can be written, using expressions (30) and (33) as

$$\langle P_2\rangle = \frac{1}{N(N-1)} \sum_{K=1}^{N} \Pi_K \sum_{j=1}^{N} j^2 \langle M_j\rangle_{N,K} - \frac{1}{N-1}. \quad (34)$$

This approach can be generalized to the probability of having $n \geq 2$ particles together.

3 Examples of coagulation-fragmentation with a finite number of particles

We shall now summarize several results in the three examples:

1. $a_i = a$
2. $a_i = a$ for $i < M$ and $a_i = 0$ if $i \geq M$
3. We finally consider the case $a_i = ai$.

3.1 Example 1: the case $a_i = a$

When $a_i = a$, the separation and formation rates s_K and f_K are computed with $F(i,j) = a$ and $C(i,j) = 1$. A cluster of size n dissociates at a rate $\sum_{i=1}^{n-1} F(i, n-i) = (n-1)a$ and the sizes of the resulting clusters are uniformly distributed between 1 and $n-1$. The total transition rate from a configuration of K to $K+1$ clusters is the sum over all possible dissociation rates

$$s_K = \sum_{i=1}^{K} (n_i - 1)a = (N-K)a. \tag{35}$$

The formation rate is proportional to the number of pairs

$$f_K = \frac{K(K-1)}{2}. \tag{36}$$

The steady-state probability Π_K for the number of clusters of size K satisfies the time independent master equation

$$\begin{cases} s_1 \Pi_1 & = & f_2 \Pi_2, \\ \mu_1 (f_K + s_K) \Pi_K & = & f_{K+1} \Pi_{K+1} + s_{K-1} \Pi_{K-1}, \\ f_N \Pi_N & = & s_{N-1} \Pi_{N-1}, \end{cases} \tag{37}$$

which leads to the relation

$$\Pi_{K+1} = (2a)^K \frac{(N-1)!}{K!(K+1)!(N-K-1)!} \Pi_1. \tag{38}$$

With he normalization condition $\sum_K \Pi_K = 1$, the probability Π_1 is expressed with a hypergeometric series

$$\Pi_1 = \frac{1}{{}_1F_1(-N+1; 2; -2a)}, \tag{39}$$

where

$${}_1F_1(a; b; z) = \sum_{n=0}^{\infty} \frac{(a)_n}{(b)_n} \frac{z^n}{n!}, \tag{40}$$

is Kummer's confluent hypergeometric function ([1] pp. 503–535) and

$$(x)_n = x(x+1)...(x+n-1) \tag{41}$$

is the Pochhammer symbol. The average number of clusters at steady state

$$\begin{aligned}\mu_1(a) &= \sum_{K=1}^{N} K\Pi_K \\ &= \Pi_1 \frac{d}{dz}\Big(z\,{}_1F_1(-N+1;2;z)\Big)_{|z=-2a}. \end{aligned} \tag{42}$$

The derivative of the Kummer's function is

$$\frac{d}{dz}\,{}_1F_1(a;b;z) = \frac{a}{b}\,{}_1F_1(a+1;b+1;z). \tag{43}$$

The mean number of clusters is expressed as

$$\begin{aligned}\mu_1(a) &= 1 + a(N-1)\frac{{}_1F_1(-N+2;3;-2a)}{{}_1F_1(-N+1;2;-2a)}, \\ &= 1 + a(N-1)G_1, \end{aligned} \tag{44}$$

where we note G_1 the function defined by

$$G_1 = \frac{{}_1F_1(-N+2;3;-2a)}{{}_1F_1(-N+1;2;-2a)}. \tag{45}$$

More generally, we introduce the functions G_i defined by

$$G_i = \frac{{}_1F_1(-N+1+i;2+i;-2a)}{{}_1F_1(-N+1;2;-2a)}. \tag{46}$$

All moments of the probability distribution Π_K can be computed and the n^{th}-order moment μ_n is expressed using the operator H defined by

$$H(f)(z) = \frac{d}{dz} z f(z), \tag{47}$$

by

$$\mu_n = \sum_{n=1}^{N} K^n \Pi_K = \frac{H^{(n)}({}_1F_1(-N+1;2;z))_{|z=-2a}}{{}_1F_1(-N+1;2;-2a)}. \tag{48}$$

Using the differentiation formula for the hypergeometric function (43), the moments are

$$\mu_n = \sum_{k=0}^{n} \alpha_k^n \frac{\Pi_{k+1}}{\Pi_1} G_k, \tag{49}$$

where

$$\alpha_k^n = \begin{cases} k! \sum_{j=0}^{k/2} (-1)^j \dfrac{(k+1-j)^n + (j+1)^n}{(k-j)!} & \text{if } k \text{ is even,} \\ k! \sum_{j=0}^{(k-1)/2} (-1)^j \dfrac{(k+1-j)^n - (j+1)^n}{(k-j)!} & \text{if } k \text{ is odd,} \end{cases}$$

and $\alpha_0^n = \alpha_n^n = 1$. The variance of the number of clusters is given by

$$\langle V_\infty(a) \rangle = \mu_2 - \mu_1^2 = a(N-1)G_1(a,N) + \frac{2}{3}a^2(N-1)(N-2)G_2(a,N) - a^2(N-1)^2 G_1^2(a,N). \quad (50)$$

3.1.1 Number of clusters of a given size

The statistical moments for the size of clusters are computed from relation (17) and the mean number of clusters of size n when there are K clusters is

$$\langle M_n \rangle_{N,K} = \sum_{(m_i) \in P'_{N,K}} m_n p'(m_i|K) = a \frac{C_{N-n,K-1}}{C_{N,K}}. \quad (51)$$

The normalizing constant $C_{N,K}$ given in eq. (15) is the $N-$th order coefficient of S^K, where S is the generating function

$$S(x) = \sum_{i=1}^{\infty} a_i x^i = a \frac{x}{1-x}. \quad (52)$$

The coefficient $C_{N,K}$ is thus equal to the $N-K$th order coefficient of $\frac{1}{K!}\frac{a^K}{(1-x)^K}$. By differentiating $N-K$ times $\frac{1}{(1-x)^K}$ and estimating the derivative at $x=0$. We obtain that

$$C_{N,K} = \frac{a^K}{K!} \frac{(N-1)!}{(K-1)!(N-K)!}. \quad (53)$$

Thus, by combining (51) and (53),

$$\langle M_n \rangle_{N,K} = \frac{(N-n-1)!K!(N-K)!}{(N-1)!(K-2)!(N-n-K+1)!}, \quad (54)$$

The mean number of clusters of size n is obtained by summing over all possibilities configuration with K clusters,

$$\langle M_n \rangle = \sum_{K=1}^{N} \langle M_n \rangle_{N,K} \Pi_K = \frac{(N-n-1)!}{(N-1)!} \sum_K \frac{K(K-1)(N-K)!}{(N-n-K+1)!} \Pi_K.$$

Using expression (38) for Π_K, we obtain

$$\langle M_n \rangle = 2a \frac{{}_1F_1(-N+1+n;2;-2a)}{{}_1F_1(-N+1;2;-2a)} \text{ if } n < N, \tag{55}$$

and

$$\langle M_N \rangle = \frac{1}{{}_1F_1(-N+1;2;-2a)}. \tag{56}$$

The mean number of clusters of size N is exactly equal to the probability $\Pi_1(N)$ of having one cluster when there is N particles (see eq. (39)).

3.1.2 Probability to find two particles in the same cluster

The probability to find two particles in the same cluster for a constant kernel $a_i = a$, when there are N particles, is

$$\langle P_2 \rangle = G_1, \tag{57}$$

where G_1 is defined in (45). Indeed the probability that two particles are in the same cluster is

$$\begin{aligned} \langle P_2 \rangle &= \frac{1}{N(N-1)} \sum_{K=1}^{N} \Pi_K \sum_{j=1}^{N} j^2 \langle M_j \rangle_{N,K} - \frac{1}{N-1} \\ &= \frac{1}{N(N-1)} \sum_{K=1}^{N} \Pi_K \left(N + 2N \frac{N-K}{K+1} \right) - \frac{1}{N-1}, \end{aligned} \tag{58}$$

where the average number of clusters of size j when there is a total of K clusters is given by relation (54). Thus,

$$\langle P_2 \rangle = \frac{2}{N-1} \sum_{K=1}^{N} \Pi_K \frac{N-K}{K+1} = -\frac{2}{N-1} + 2\frac{N+1}{N-1} \sum_{K=1}^{N} \frac{1}{K+1} \Pi_K \tag{59}$$

which is the definition of G_1 eq. (45). For large N, we thus obtain that the probability that two particles are in the same cluster is

$$\langle P_2 \rangle \approx \sqrt{\frac{2}{aN}}. \tag{60}$$

The results presented in this section were used to study the distribution of clusters in biological systems such as telomere organization in yeast [17].

3.2 Example 2: the case $a_i = a$ for $i < M$ and $a_i = 0$ if $i \geq M$

When N particles can associate or dissociate with a constant rate, but cannot form clusters of more than M particles, the configuration space for the distribution of N particles in K clusters of size less than M is now

$$P'_{N,K,M} = \{(m_i)_{1\leq i\leq M}; \sum_{i=1}^{M} i m_i = N,\ \sum_{i=1}^{M} m_i = K\}. \tag{61}$$

First, the minimal number of clusters is necessarily bounded by $K \geq N/M$, since the opposite would imply a cluster of at least $M+1$ particles. The probability of a configuration $(m_1, ..., m_M) \in P'_{N,K,M}$ is equal to

$$P'\{(m_1, ..., m_M) \in P'_{N,K,M}\} = \frac{1}{C_{N,K,M}} \frac{1}{m_1!...m_M!}, \tag{62}$$

where the normalization constant $C_{N,K,M}$ is the Nth order coefficient of

$$(aX + aX^2 + ... + aX^M)^K = a^K \frac{1}{(1-X)^K} \sum_{n=0}^{K} \binom{K}{n} (-1)^n X^{nM+K}. \tag{63}$$

Then the Nth order coefficient of the polynomial is obtained by finding the $(N - nM - K)$th order coefficient of $(1-X)^{-K}$

$$C_{N,K,M} = a^K \sum_{n=0}^{K} \binom{K}{n} (-1)^n \frac{1}{(N-(nM+K))!} D^{(N-(nM+K))} \left(\frac{1}{(1-X)^K} \right)_{|X=0}, \tag{64}$$

where we write $D^{(n)}$ the $n-$th order derivative. Thus, setting $K_0 = \lfloor \frac{N-K}{M} \rfloor$, where $\lfloor . \rfloor$ is the floor function, we have

$$C_{N,K,M} = a^K K \sum_{n=0}^{K_0} \frac{(N-nM-1)!}{n!(K-n)!(N-(nM+K))!} (-1)^n. \tag{65}$$

For $M = N$ we find $K_0 = 0$ and the normalization constant

$$C_{N,K,N} = a^K \frac{(N-1)!}{(K-1)!(N-K)!}, \tag{66}$$

is equal to the normalization constant $C_{N,K}$ obtained for the constant kernel in section 3.1. The mean number of clusters of size $i \leq M$ conditioned on the number of clusters K is

$$\langle M_i \rangle_K = \sum_{m_i \in P'_{N,K,M}} m_i p'(m_1, ...m_M) = a \frac{C_{N-i,K-1,M}}{C_{N,K,M}}. \tag{67}$$

Two clusters of size i and j can form a new cluster only if $i+j \leq M$. The formation rate when there are K clusters is thus

$$f_K = \sum_{(m_i)\in P'_{N,K,M}} p'(m_1,\ldots,m_N)\left(\sum_{i=1}^{M/2}\frac{m_i(m_i-1)}{2} + \sum_{\substack{i,j=1\\ i+j\leq M; i\neq j}}^{M} m_i m_j\right). \tag{68}$$

The formation rate can be written as a function of the coefficients $C_{N,K,M}$ as

$$f_2 = C_{N,2,M}, \tag{69}$$

and for $K > 2$

$$\begin{aligned} f_K &= \frac{K(K-1)}{2}\sum_{i=1}^{\min(\frac{M}{2},\frac{N-K+2}{2})} C_{N-2i,K-2,M} \\ &+ \frac{K(K-1)}{2}\sum_{\substack{i,j=1\\ i+j\leq M}}^{\min(M-1,N-K+1)} C_{N-i-j,K-2,M}. \end{aligned} \tag{70}$$

The separation rate remains unchanged $s_K = (N-K)a$, and the probabilities at steady state are given by

$$\Pi_K = \frac{f_{K+1}}{s_K}\Pi_{K+1}. \tag{71}$$

We illustrate the limit case $a \to 0$ for $N = 9$, $M = 4$ (Fig. 2). When $a > 0$, all partitions are accessible, but as $a \to 0$, the steady state configurations are dominated by the configurations with the largest possible cluster size $(4,4,1)$, $(4,3,2)$ and $(3,3,3)$. Applying formulas (65) and (67), we obtain the limit cluster configuration probabilities

$$\begin{aligned} p(4,4,1) &= \frac{3}{10} \\ p(4,3,2) &= \frac{6}{10} \\ p(3,3,3) &= \frac{1}{10}. \end{aligned} \tag{72}$$

These steady state probabilities do not depend on the initial particles configurations as long as $a \neq 0$. For $a = 0$, there are three possible configurations

$(4,4,1)$, $(4,3,2)$ and $(3,3,3)$: once equilibrium is attained, the clusters will remain unchanged. The probability to get to equilibrium depends on the configuration and the order of clustering events. When there is no limitation in the cluster formation ($M = N = 9$), a single cluster containing all particles is formed (Fig. 2, left panel). For large values of a, most clusters are very small, and the distributions are similar for $M = 4$ and $M = 9$ (Fig. 2, right panel).

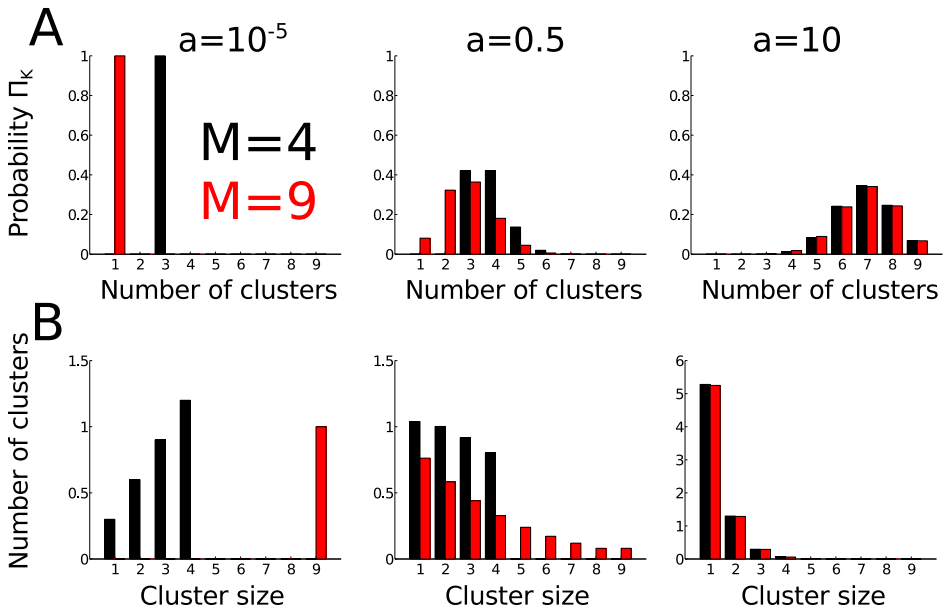

Figure 2: (**A**) Distribution of the number of clusters Π_K for $N = 9$, when cluster sizes are limited ($M = 4$, black) and not limited ($M = 9$, red). There is a minimum of $\lceil N/M \rceil$ clusters. From left to right : $a = 10^{-5}$, $a = 0.5$, $a = 10$. (**B**) Mean number of clusters of each size $\langle M_n \rangle$. For $a \to 0$, for $N = 9$ and $M = 4$ the clusters organize in three different cluster configurations, while for $M = N$ a single cluster containing N particles is formed.

The probability for two particles to be in the same cluster provides a good estimation for the cluster distribution for various values of the parameter a (Fig. 3). When a is large, most particles are contained in very small clusters and the probability $\langle P_2 \rangle$ is similar for the cases $M = 4$ and $M = 9$. When $a \to 0$, particles tend to form larger clusters. A single cluster containing all particles is formed and $\langle P_2 \rangle \to 1$ when $M = 9$, but the maximal value

of $\langle P_2 \rangle$ is less than 1 when the maximal cluster size is limited. We can explicitly compute $\langle P_2 \rangle$ in the limit case $a \to 0$. For example for $M = 4$, using eq. (30), and summing over all possible configurations (72), we obtain

$$\begin{aligned} \langle P_2 \rangle &= p(4,4,1)P_2(4,4,1) + p(4,3,2)P_2(4,3,2) + p(3,3,3)P_2(3,3,3) \\ &= \frac{3}{10}\frac{24}{72} + \frac{6}{10}\frac{20}{72} + \frac{1}{10}\frac{18}{72} \\ &= \frac{7}{24}. \end{aligned}$$

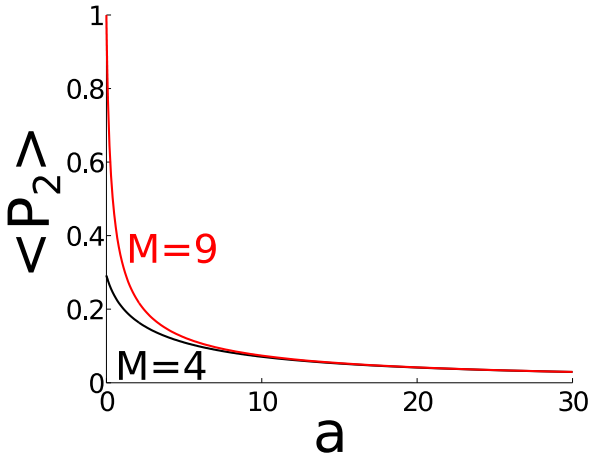

Figure 3: **Probability $\langle P_2 \rangle$ that two particles are in the same cluster.** The parameters are $N = 9$ and $M = 4$ (black), $M = 9$ (red). For large values of $a \gg 1$, only small clusters are present and the steady state distributions are similar for the cases $M = 4$ and $M = 9$. When $a \to 0$ the clusters organize in three different cluster configurations, while for $M = N$ a single cluster containing N particles is formed.

3.3 Example 3: Application to the case $a_i = ai$

We consider the case $a_i = ai$. The number of clusters of size i is asymptotically [13]

$$\langle M_i \rangle = aie^{-i\sqrt{2a/N}}. \tag{73}$$

Similarly to the previous examples, the number of clusters of size i, for a given distribution of K clusters is

$$\langle M_i \rangle_{N,K} = i \frac{\binom{N-i+K-2}{N-K-i+1}}{\binom{N+K-1}{N-K}}. \tag{74}$$

The number of clusters of a given size is determined by the probability of a distribution of K clusters Π_K. It is given in the induction relation

$$\Pi_K = \frac{f_{K+1}}{s_K} \Pi_{K+1}, \tag{75}$$

where f and s are the formation and separation rates. The coagulation kernel is $C(i,j) = 1$ and the fragmentation kernel $F(i,j) = a\frac{ij}{i+j}$, and we obtain that

$$d(n) = \sum_{i=1}^{n-1} a \frac{i(n-i)}{n} = \frac{a(n^2-1)}{6}. \tag{76}$$

The separation rates are

$$s_1 = \frac{a(N^2-1)}{6} \tag{77}$$

and for $K \geq 2$

$$s_K = \frac{a}{6} \frac{1}{\binom{N+K-1}{N-K}} \frac{1}{(2K-3)!} \sum_{i=1}^{N-K+1} \frac{i(i^2-1)(N-i+K-2)!}{(N-i-K+1)!}. \tag{78}$$

In addition,

$$\begin{aligned} s_K &= \frac{a}{6} \frac{(2K-1)(N+K-2)}{N+K-1} ((N+K-2)^2+5) \\ &- a(2N+2K-3)(2K-2) + \frac{a}{2}(N+K)^2 \frac{(2K-2)(2K-1)}{2K} \\ &- \frac{a}{6}(N+K)(N+K+1) \frac{(2K-1)(2K-2)}{2K+1}. \end{aligned} \tag{79}$$

The formation rates are obtained from the number of cluster pairs that can coagulate, and are given by

$$f_K = \frac{K(K-1)}{2}. \tag{80}$$

To conclude this section, model of discrete coagulation-fragmentation processes with a finite number of particles are used to determine the steady state probability distribution when the number of clusters is fixed. Using the partitions of the total number of particles with a given number of clusters, various statistical quantities and moments such as the cluster distributions can be computed, including also the mean number of clusters of a given size conditioned on the total number of clusters. Two times can be used to characterize the cluster dynamics: one is the time that two particles spend together and second is the time they spend separated. In the next section, we will describe specific applications in cell biology.

4 Modeling and simulations of telomere coagulation-fragmentation process

Telomere aggregate and dissociate according to the coagulation-fragmentation process presented in section 3.1. To obtain numerically any quantity of interest, we use the master equations that describe the coagulation-fragmentation process. The equation that describes the probability $P(n_1, \ldots, n_N, t)$ of having a distribution of N clusters distributed in clusters of size n_1 , ..., n_N, given in eq. (6), is t

$$\begin{aligned}
\frac{d}{dt}P(n_1,\ldots,n_N,t) &= -\left(\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} C(n_i,n_j) + \sum_{i=1}^{N}\sum_{k=1}^{n_i-1} F(k,n_i-k)\right) P(n_1,\ldots,n_N,t) \\
&+ \sum_{k=1}^{N} \sum_{\substack{n_i'>0, n_j'>0 \\ n_i'+n_j'=n_k}} C(n_i',n_j') P(n_1,\ldots,n_i',\ldots,n_j',\ldots,n_N,t) \\
&+ \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} F(n_i,n_j) P(n_1,\ldots,n_i+n_j,\ldots,n_N,t), \qquad (81)
\end{aligned}$$

where $C(i,j) = k_f$ is the formation rate of a cluster of size $i+j$ from two clusters of sizes i and j, and $F(i,j) = k_b$ is the rate of dissociation of a cluster of size $i+j$ into two clusters of size i and j.

Because the time distribution of the telomere to a small target is exponential, the encounter rate of telomeres at the nuclear periphery can be characterized by a single parameter (the arrival rate or equivalently by an effective diffusion constant). Even though telomere motion involves complex polymer chains accounting for the physical chromosomal chain, any encounter is a rare event, and its rate is Poissonian. Consequently, to model clustering, we use this property to approximate the arrival time of a chromosome to a small cluster by the Poissonian dynamics, as long as the chromosome length does not restrict the motion of the telomere on the nuclear surface. Two telomeres encounter at a Poissonian rate k_f.

Polymer simulations (Fig. 4B) confirms that the arrival time of a telomere to a cluster can be simulated using a Poissonian distribution approach. In that case, it is enough to study the dynamics of 32 stochastic particles (Fig. 4C). Thus, using a molecular dynamics simulation of two Brownian particles on the surface of a sphere [9], Brownian simulations of particle located on the two-dimensional sphere except for a region of the size of the

nucleolus (see earlier discussion) leads to an approximation for the forward rate of $k_f \approx 1.9\ 10^{-3} s^{-1}$, where the encounter disk is of radius $\delta = 0.015 \mu m$ and the effective diffusion constant is $D = 0.005 \mu m^2/s$ [8].

When a telomere aggregates to a cluster, it only slightly varies in size. Indeed, in the complex environment of the nuclear surface, the diffusion constant varies with the log of the radius of the effective diffusing particle. Thus any changes in the radius will result only in a small change in the diffusion coefficient. We neglected any possible changes in the scattering cross section and motility, which could modify the forward binding rate [17]. Thus the encounter rate between clusters or telomeres will be approximated by a constant independent of the size.

In the Gillespie algorithm, the transition rate constants between different cluster configurations are given as follows: for a distribution $(n_1, \ldots, n_K)$ of clusters, the transition probabilities to the neighboring states depend on two events: either two clusters (n_i, n_j) associate to form a new cluster of size $n_i' = n_i + n_j$ with an association rate k_f or a cluster of size n dissociates into two, with a rate $(n-1)k_b$ that depends on the number of bonds. The size of the resulting dissociated clusters is uniformly distributed in the interval $[1, n-1]$. Since there are $\frac{K(K-1)}{2}$ pairs, the association rate equals $\frac{K(K-1)}{2} k_f$, and the total fragmentation rate is the sum over all dissociation rates $\sum_j (n_j - 1) k_b = (N-K) k_b$. The total transition rate from the state $(n_1, \ldots, n_K)$ to any of the possible association and dissociation events is $a_0(n_1, \ldots, n_K) = \sum a_i = \frac{K(K-1)}{2} k_f + (N-K) k_b$. Each iteration step of the algorithm uses the classical Poissonian random transition time $\tau = -\frac{\log r_1}{a_0}$, where r_1 is a uniform random variable in $[0, 1]$ and each reaction event i has a probability $\frac{a_i}{a_0}$ to occur, and the chosen reaction i is sorted out using the criteria $\sum_{j=1}^{i-1} \frac{a_j}{a_0} < u \leq \sum_{j=1}^{i} \frac{a_j}{a_0}$ where u is uniformly distributed in $[0, 1]$.

4.1 Influence of the chromosome arm length on the clustering dynamics

Because chromosome arms with a length below 300 kb are mainly located in a small region near the spindle pole body (SPB) [47], while telomeres of longer chromosome arms exhibit motion near the nucleolus, we decided to integrate these constraints into the telomere dynamics (Fig. 4D). We distribute telomeres into two classes based on the length of the chromosome arm [47] and restricted 12 telomeres to a small region account for short-short interactions (SS) around the SPB (1/3 of the surface) and the other 20 are free to diffuse in a larger region where only long telomeres can interact (LL),

which excludes both the nucleolus and a small cap around the SPB (SL is 2/3 of the nucleus surface).

In the common region SL, both types of telomeres can meet to form mixed clusters. There are three possible classes of telomere clusters: clusters containing telomeres from long chromosome arms only (long), from short chromosome arms only (short) or from long and short chromosome arms (mixed), leading to six forward rates, accounting for the long-long, short-short, long-short, long-mixed, short-mixed and mixed-mixed interactions.

In addition, for two telomeres from the pool of long chromosome arms, the recurrence time and we report that $T_R = 442$ s ($n = 1,000$), shorter than the forward time $k_f^{-1}(L, L) \approx 500$s. Thus, the interaction of telomeres from short chromosome arms with a cluster made of long ones will contribute to the confinement of the cluster to a smaller region of the nuclear periphery, which will consequently decrease the mean time for two telomeres to meet again. The mean time to separation T_S was similar for telomeres from short-short, short-long and long-long chromosome arms (≈ 21 s, versus 31 s for the dissociation time between two telomeres, $n = 1,000$), reflecting that clusters contain the same number of telomeres independently of their composition.

Finally, the equilibrium probability to find a given telomere in a visible cluster (containing more than 2) was $Pr(S, S) = 0.06$, $Pr(L, L) = 0.045$ and $Pr(L, S) = 0.04$ (for short-short, long-long and long-short arm interactions), confirming that the encounter rate for small telomeres is higher than for long ones, due to the smaller space they can explore. Our results are mainly consistent with [47], where the probabilities for two telomeres to belong to the same focus are determined experimentally to be mostly in the range 0.04-0.09. The differences between these experimental data and our simulations might be due to specific interactions between telomere pairs, which we did not take into account. Indeed, contacts between telomeres on opposite chromatid arms of equal length is favored [48].

The aggregation-dissociation model for telomere organization was used to extract invivo parameters by comparing stochastic simulations with live cell imaging data (Fig. 5A). The dissociation rate kb is estimated by comparing the experimental and simulation histograms for the number of clusters containing more than two telomeres (Fig. 5B). Histograms similarity was evaluated using the Kolmogorov-Smirnov (KS) score, here defined as the maximum of the absolute difference of the experimental and simulated cumulative distribution function for the number of clusters. The optimal value of the KS score was 0.11 obtained for $k_b = 2.410^{-2}s^{-1}$.

However, a higher variance in the histogram of the experimental number of clusters. To account for this variation, we introduced fluctuations in

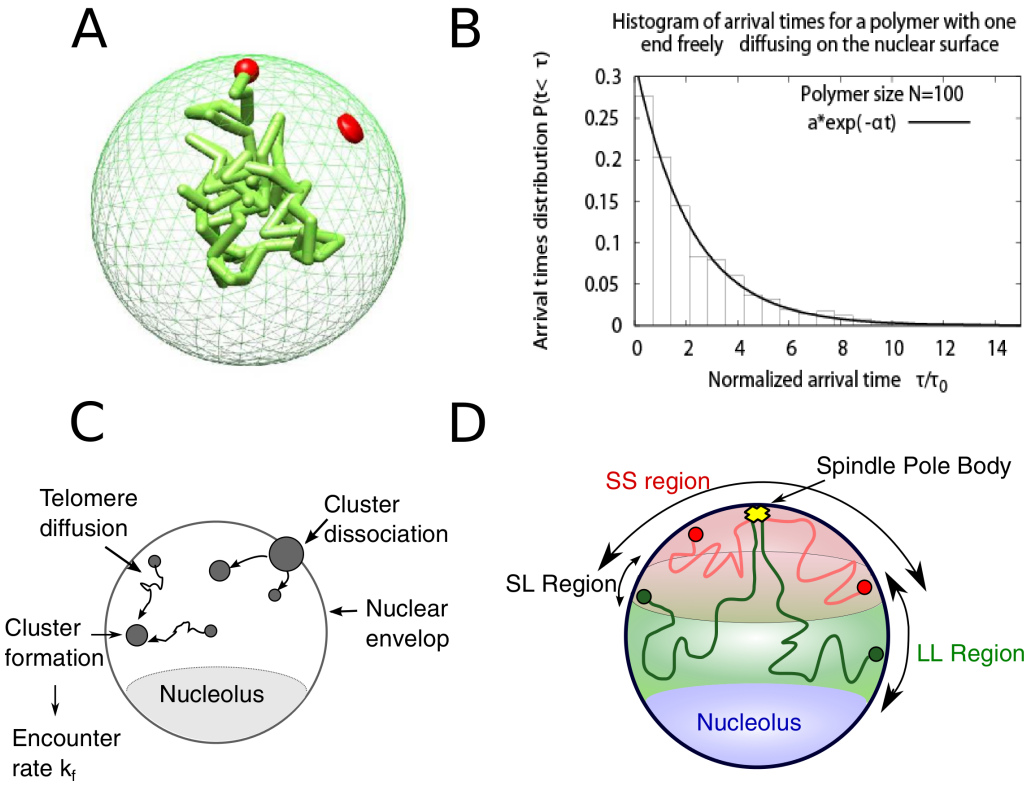

Figure 4: Computational model of telomere cluster formation. (A) Snapshot from a Brownian dynamics simulation of a polymer with one end anchored on the nuclear surface. The polymer is composed of 100 monomers with average distance between monomers of $l_0 = 50$ nm, and the nucleus is a reflecting sphere of size $R = 250$ nm. (B) Histogram of the arrival times for a polymer of 100 monomers freely diffusing in the nucleus and one end constrained to diffusion on the surface. A fit of the form $f(t) = a\exp(-bt)$ gives $a = 1.014$ and $b = 0.76$. (C) The diffusion-aggregation-dissociation model of telomere organization. Telomeres are simplified as Brownian particles diffusing on the nuclear surface that can meet and form clusters, and clusters of n telomeres split at a rate $(n-1)k_b$. The coarse-grained association rate k_f is taken as the average of the cluster meeting times. (D) Influence of long and short chromosome arms on clustering. Decomposition of the nucleus in subdomains with telomeres from short and long chromosome arms. Both types of telomere can interact in a common region.

the value of the dissociation rate k_b of each cell. We generated random values of k_b following a Gaussian distribution and we found that for $k_b = 2.310^{-2} \pm 1.310^{-2} s^{-1}$, which corresponds to $a = k_b/k_f = 12 \pm 7$, we obtain an optimal fit for the distribution of the number of clusters. Simulations show an excellent adequacy to the experimental cluster distribution (Fig. 5B), size (Fig. 5C) and size distribution (Fig. 5D), with a KS score of 0.07.

We observed an average of three detectable clusters per cell, and very few cells with more than eight clusters. Interestingly, in the simulations, 9.9 (8.2) telomeres are isolated (in pairs). In addition, the number of telomeres per cluster obtained in our simulations reflects very well the cluster intensity obtained experimentally: in both simulated and experimental data, we found that the average cluster intensity does not vary with the number of clusters per cell (Fig. 5C). Because there are 32 telomeres and that the intensity is an increasing function of the number of telomeres, we conclude that there are in average no more than four telomeres per cluster. A better precision about the cluster distribution is obtained by plotting the distribution of the first three brightest clusters for both experimental and simulated data (Fig. 5D): in both cases the three brightest clusters contain four telomeres.

The robustness of the aggregation-dissociation model is tested for the organization of telomeres in diploid cells where the nuclear volume (nucleus radius = 1.25 μm) and the number of telomeres are doubled. These changes in the cell geometry affect the forward rate, which we recomputed from Brownian simulations, and we now found for the association rate $k_f = 1.110^{-3} s^{-1}$. Considering that the backward rate is unchanged and taking the value found in the normal case, we obtained for the new equilibrium constant the value $a = 21 \pm 12$ (compared to 12 ± 7 for the haploid).

Telomere foci in diploid cells are shown in Fig. 5E, and the number of telomere foci obtained by simulation is similar to the number measured in live cells. They have in average 6 clusters containing 3 to 6 telomeres per cell (Fig. 5F,G). The light intensity and the telomeres distribution of measured and simulated telomeres per cluster were very similar (Fig. 5F-H). Interestingly the median cluster size is 4 in both haploid and diploid cells, i.e. there are four telomeres per cluster, suggesting that the number of telomeres per cell does not influence the number of telomeres per cluster. Furthermore, according to the simulations, in diploid cells, telomeres cluster in 5-9 foci containing 3 to 6 telomeres, while 18.7 telomeres are single and 16.4 are in pairs. The matching between experimental data and numerical simulations confirms the robustness of the model to parameter changes, while the physical properties of the telomeres and the cluster dissociation rate were maintained fixed.

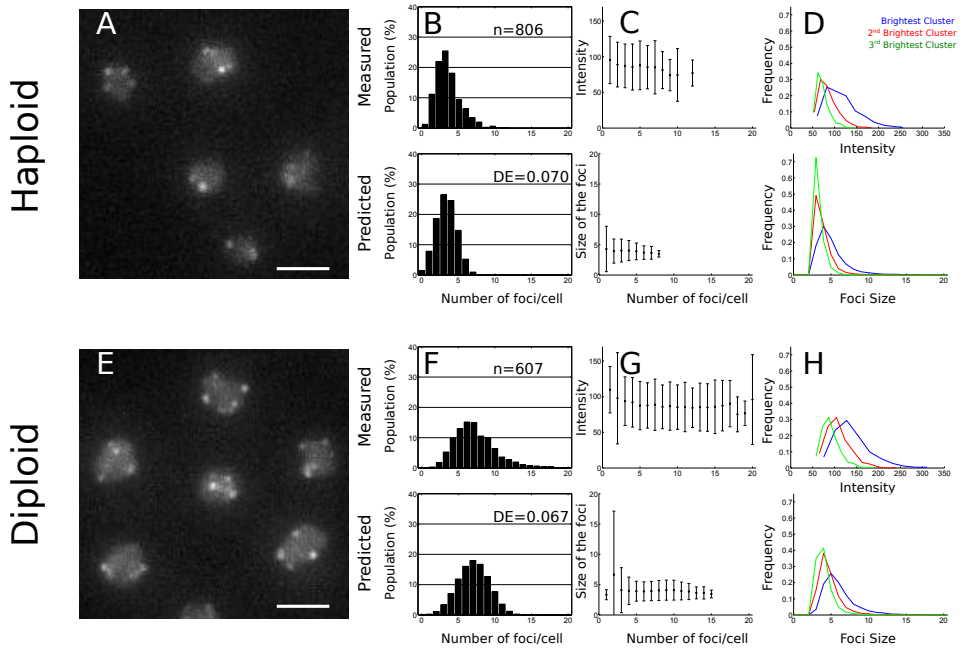

Figure 5: Comparison of experimental and simulation results of telomeres clustering in yeast. (A,E) Live-cell imaging of telomere clusters. Representative fluorescence image of the telomere-associated protein Rap1 tagged with GFP (scale bar, 2 μm) in haploid (A) and diploid cells (B). (B, F) Histogram of the number of clusters per cell. (C, G) Mean $\pm$ s.d. of the intensity distributions of the clusters in live cells and distribution of the cluster size in the Brownian simulations. In the haploid cells, clusters are made of four telomeres, with a small dispersion that does not depend on the cluster number. (D, H) Fluorescence intensity (experiments) and sizes defined as the number of telomeres per cluster (simulations) for the three brightest clusters. The frequency of occurrence (y-axis) of a given cluster size is plotted as a function of the intensity of a cluster (x-axis), proportional to the telomere number.

5 Modeling capsid formation as an aggregation process

In this second part, we study here the kinetics of cluster formation starting with the arrival and fusion of elementary particles at a nucleation site. Particles involved in the cluster formation are organized in aggregates. The aggregates increase the cluster size by fusion to the particles. This is an elementary model of capsid viral assembly, where the density distribution of aggregates is at steady state. To maintain this distribution, the protein production must be much larger than the aggregates needed to form a single capsid.

In that model, a cluster can accept a maximum of N_0 particles, and is complete when exactly N_0 particles have arrived. The cluster is formed upon the arrival of aggregates of various size. When a cluster has reached a size n, it can accept aggregates of size less than $N_0 - n$ (Fig. 6C). Each aggregate binds to a cluster with a Poissonian arrival rates λ, independent of the aggregate size. Aggregates participating in the cluster formation are already formed and are at steady state. Therefore the number n_k of aggregates containing k particles is constant. The total number of particles N_T is distributed among the aggregates., therefore

$$N_T = \sum_{k=1}^{N_0} k n_k. \tag{82}$$

We assume that the number of aggregates of size k is distributed exponentially and given by

$$n_{k+1} = p n_k, k \geq 0 \tag{83}$$

where the parameter $0 \leq p \leq 1$.We present here the models of nucleation using a mean-field approximation and a stochastic jump process.

5.1 Mean-field approximation

We now derive an equation for the cluster size $n(t)$ at time t. The cluster growth rate depends on the arrival of an aggregate of size k and on the probability q of finding a free site at the cluster. We neglected here the geometrical organization of an aggregate and consider that upon fusion, it fills empty slots in the cluster. We do account here for the geometrical organization in facet of aggregates which participate to the structure of viral capsids. Thus, the probability q does not depends on the geometry or

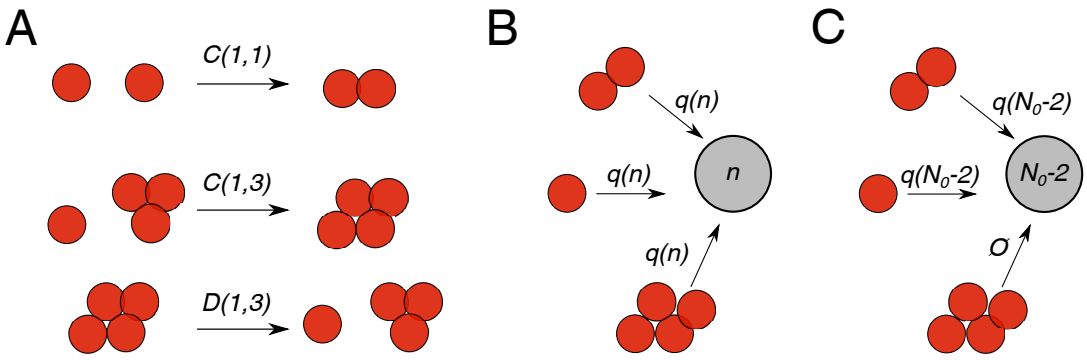

Figure 6: Schematics of clustering for a finite number of particles. (A) Model used for telomere clusters coagulation-fragmentation. (B,C) Model used for capsid formation.

positions of the aggregates already present in the cluster but only on their number. We chose the linear relation

$$q(n(t)) = 1 - \frac{n(t)}{N_0}. \tag{84}$$

In addition, we neglected any changes of the arrival rate due to the size of the aggregate that can affect the diffusion coefficient. To conclude, a nucleation site is formed when it is entirely filled by aggregates. The cluster growth is due to the arrival of aggregate of size k and the rate is $\lambda k n_k$. The cluster total growth rate is is the product of the probability to find an available site times the sum of the arrival rate of any aggregate. The average size $n(t)$ satisfies the equation

$$\dot{n}(t) \quad = \quad \lambda \left(1 - \frac{n(t)}{N_0}\right) \left(\sum_{k=1}^{N_0 - n(t)} k n_k\right), \tag{85}$$

which reduces to

$$\dot{n}(t) \quad = \quad A(N_0 - n(t)) \left[1 - p^{N_0 - n(t)}(1 + (N_0 - n(t))(1-p))\right], \tag{86}$$

with initial condition $n(0) = 0$ and

$$A = \frac{\lambda N_T}{N_0 \left[1 - p^{N_0}(1 + N_0(1-p))\right]}. \tag{87}$$

For $0 < p < 1$, although eq. (86) cannot be integrated analytically, we obtain the short and long time asymptotic in the limit of N_0 large. For short time, the size for the growing cluster is

$$n(t) \approx \lambda N_T t, \text{ for } t \ll 1, \tag{88}$$

which is independent of p. For large t and small p, the first order expansion is

$$n(t) \approx N_0 - \frac{N_0}{\lambda N_T (p - 1 - \log p)} \frac{1}{t} \text{ for } t \gg 1. \tag{89}$$

In the limit case $p = 1$, equation (86) changes its nature and for large N_0, it reduces to

$$n(t) = N_0 - \frac{N_0}{\sqrt{1 + 2\lambda \frac{N_T}{N_0} t}} \text{ for } t \gg 1. \tag{90}$$

When there are single monomers only ($p = 0$) eq. (85) describes the classical kinetics of arrival and the solution to eq. (86), which reduces to a single exponential $n(t) = N_0(1 - e^{-\lambda \frac{N_T}{N_0} t})$. We plotted in figure 7A the kinetics of the cluster formation. Interestingly, the cluster is formed more quickly for lower values of the parameter p.

5.2 A stochastic dynamics for the cluster formation

Due to the discrete arrival of aggregates to the nucleation site, the cluster size increases by random jumps that we shall describe now using a stochastic jump process. When an aggregate arrives in the time interval $(t, t + \Delta t)$, the cluster of size $n(t)$ at time t increases with a probability $\mu(n(t))dt$ that depends on its size at time t, thus

$$n(t + \Delta t) = \begin{cases} n(t) & \text{w.p. } 1 - \mu(n(t))\Delta t \\ \\ n(t) + J(n(t)) & \text{w.p. } \mu(n(t))\Delta t, \end{cases}$$

where $J(n(t))$ is the size of a random jump, characterized by its conditional transition distribution

$$\begin{aligned} Pr\{J(n(t)) = m - n | n(t) = n\} &= w(m - n|n) \\ &= \frac{(1-p)p^{m-n}}{p(1 - p^{N_0 - n})} \end{aligned}$$

and $w(m-n|n)$ is transition probability from n to m, which we normalized by summing over all aggregate sizes that can fuse with the cluster. To determine the arrival rate of an aggregate, we start with a cluster containing n particles. The arrival rate of an aggregate is given by the jump rate $\mu(n)$, which is equal to the arrival rate λ of an aggregate of particles (or a single particle), multiplied by the number of aggregates smaller than $N_0 - n$, so they can enter in the nucleation site, multiplied by the probability of finding a free site (proportional to $1 - \frac{n}{N_0}$). The jump rate is thus

$$\begin{aligned} \mu(n) &= \lambda(1 - \frac{n}{N_0}) \sum_{k=1}^{N_0-n} n_k \\ &= a(N_0 - n)(1 - p^{N_0-n}), \end{aligned} \tag{91}$$

where $a = \lambda \frac{N_T}{N_0} \frac{1-p}{1-p^{N_0}(1+N_0(1-p))}$. The probability density function satisfies the master equation

$$\begin{aligned} p(m, t+\Delta t) &= (1 - \mu(m)\Delta t)p(m,t) \\ &+ \sum_{n=1}^{m-1} w(m-n|n)p(n,t)\mu(n)\Delta t, \end{aligned}$$

which tends in the limit $\Delta t \to 0$, to the discrete forward Fokker-Planck equation

$$\begin{aligned} \frac{\partial p(m,t)}{\partial t} &= L_m p(m,t) = -\mu(m)p(m,t) \\ &+ \sum_{n=1}^{m-1} \mu(n)w(m-n|n)p(n,t), \end{aligned} \tag{92}$$

where L_m is the forward Kolmogorov operator.

5.3 The mean time to cluster formation.

The time to form the cluster is the mean first passage time $\langle \tau(n) \rangle$ of the cluster size to its maximum N_0. By definition,

$$\tau(n) = \inf \{t > 0; n(t) \geq N_0 | n(0) = n\} \tag{93}$$

and the MFPT is solution of the backward equation [43] with absorbing boundary condition at N_0

$$\begin{cases} L_n^* \langle \tau(n) \rangle &= -1 \\ \\ \langle \tau(N_0) \rangle &= 0, \end{cases} \tag{94}$$

where the operator L_n^* is the adjoint of L_m. The MFPT is obtained by solving the system of equations for $0 \leq n \leq N_0 - 1$,

$$-1 = -\mu(n)\langle\tau(n)\rangle + \sum_{m=n+1}^{N_0} \langle\tau(m)\rangle\mu(n)w(m-n|n). \tag{95}$$

The mean time of a cluster formation is then

$$\begin{aligned}\langle\tau(0)\rangle &= \frac{N_0(1-p^{N_0}(1+N_0(1-p)))}{\lambda N_T} \times \\ &\left[\frac{1}{N_0(1-p)(1-p^{N_0})} + \sum_{i=1}^{N_0-1} \frac{1}{i(1-p^{i+1})(1-p^i)}\right],\end{aligned} \tag{96}$$

which depends on the total number of particles N_T, the maximal size N_0, the parameter p that describes the size distribution of aggregates and λ the arrival rate of an aggregate to the nucleation site.

For small p and large N_0, we obtain the approximation

$$\langle\tau(0)\rangle = \frac{N_0}{\lambda N_T}(\log N_0 + \gamma) + o(p). \tag{97}$$

For a large nucleation site N_0, in the limit $p \to 1$, the mean time remains finite and equation (96) becomes

$$\langle\tau(0)\rangle(N_0, p, \lambda, N_T) = \frac{N_0^2}{\lambda N_T}(\frac{\pi^2}{6} - 1) \text{ for } p = 1. \tag{98}$$

The mean formation time does increases drastically as p tends to 1 (Fig. 7C). Indeed, a cluster starts growing very rapidly when large aggregates arrive, however the growth is reduced later on because the number of admissible aggregates (smaller than the number of available sites) is small. Admissible aggregates represent only a small fraction of the total number of aggregates.

5.4 Composition of a cluster

We now characterize the cluster assembly by studying the size distribution of aggregates that have arrived to the nucleation site. We shall derive also the size of the largest aggregate that contributes to the cluster formation. To evaluate the various sizes of aggregates that bind to the cluster, we consider

the ensemble of aggregates. During the sequential steps of aggregation, the number of particles C_n at the nth-step, with $C_0 = 0$, follows the equation

$$C_{i+1} = C_i + z_{i+1}, \tag{99}$$

when z_{i+1} is the size of the aggregate that binds at step $i+1$. The cluster contains a maximum of N_0 particles. The size of aggregate z_{i+1} that binds to the cluster can take values in $(1, \ldots, N_0 - C_i)$ and thus the probability that the $i+1$th aggregate is of size k when there are $N_0 - C_i$ free sites is

$$P_{N_0-C_i}(z_{i+1} = k) = \frac{(1-p)p^{k-1}}{1-p^{N_0-C_i}}. \tag{100}$$

Thus, the joint probability that the cluster assembles with the following order of arrival $(k_1, \ldots, k_n)$ is the product of the conditional probabilities (100)

$$P(z_1 = k_1, \ldots, z_n = k_n) = \prod_{i=1}^{n} P_{N_0-\sum_{j=1}^{i-1} k_j}(z_i = k_i),$$

with the condition $\sum_{i=1}^{n} k_i = N_0$.

5.5 The largest aggregate merging to the cluster

The probability that the largest aggregate z_{max} is less than K during the cluster assembly is

$$P_{N_0,K} = \sum_{\substack{\{(k_1,..k_n);\sum k_i = N_0\} \\ \text{and } k_1,\ldots,k_n \leq K}} P(z_1 = k_1, \ldots, z_n = k_n). \tag{101}$$

To obtain an approximation of $P_{N_0,K}$ for large N_0, we sum over the first jump size, which leads to the induction formula

$$P_{N_0,K} = \sum_{k_1=1}^{K} P_{N_0}(z_1 = k_1) P_{N_0-k_1,K}, \tag{102}$$

where $P_{n,k} = 1$ for $n \leq k$. When the parameter $p = 1$, formula (102) reduces to

$$P_{N_0,K} = \sum_{k_1=1}^{K} \frac{1}{N_0} P_{N_0-k_1,K}. \tag{103}$$

The induction formula (103) can be solved by $P_{N,K} = f(\frac{K}{N})$, where f satisfies $f'(x) = \dfrac{f(\frac{x}{1-x})}{x}$ with $x = \frac{K}{N}$. The function f is solution of a n-th order linear differential equation on each interval $[\frac{1}{n+1}, \frac{1}{n}]$ for $n \geq 1$, which we solved on the intervals $(\frac{1}{3}, \frac{1}{2})$ and $(\frac{1}{2}, 1)$. However there is no simple formula on other intervals. Finally, the probability for the size of the largest aggregate to be less than K after the cluster is filled is given for large N_0 by

$$P_{N_0,K} = 1 + \log \frac{K}{N_0} \text{ for } N_0/2 \leq K \leq N_0 \tag{104}$$

and

$$\begin{aligned} P_{N_0,K} &= 1 + \log \frac{K}{N_0} + \mathrm{Li}_2(\frac{K}{N_0}) + \frac{1}{2}\log^2 \frac{K}{N_0} \\ &+ \log(\frac{K}{N_0}) + 1 \text{ for } N_0/3 \leq K \leq N_0/2 \end{aligned} \tag{105}$$

where $\mathrm{Li}_2(x) = \sum_{k=1}^{\infty} \frac{x^k}{k^2}$. The probability $P_{N_0,K}$ is well approximated by the function f that we constructed inductively on intervals $(\frac{1}{3}, \frac{1}{2})$ and $(\frac{1}{2}, 1)$ (Fig. 7D). The construction of the function f reflects that the number of possible jumps of maximal size is limited: indeed, once an aggregate of size larger than $N_0/2$ has arrived, the size of all other aggregates can only be smaller than $N_0/2$, leading to the initial interval $(\frac{1}{2}, 1)$. Similarly, after an aggregate of size between $N_0/3$ and $N_0/2$ has arrived, other possible aggregates have a size smaller than $N_0/3$. This constraint leads to the second interval $(\frac{1}{3}, \frac{1}{2})$. We obtain by induction the division in intervals $(\frac{1}{n+1}, \frac{1}{n})$.

5.6 Aggregation of capsid in potential wells

Recent super-resolution data have revealed that GAG proteins of the HIV virus can aggregate in specific microdomains [15]. Interestingly, the proteins aggregate in small regions characterized by a physical potential well (fig. 8), discovered in [17]. Indeed the motion of aggregates on the membrane surface is influenced by a diffusion coefficient D and a field of force $F(X,t)$, following the overdamped Langevin model equation

$$\dot{X} = \frac{F(X(t),t)}{\gamma} + \sqrt{2D}\,\dot{W}, \tag{106}$$

where W is a Gaussian white noise and γ is the dynamical viscosity [34]. The source of the noise is the thermal agitation of the ambient lipid and

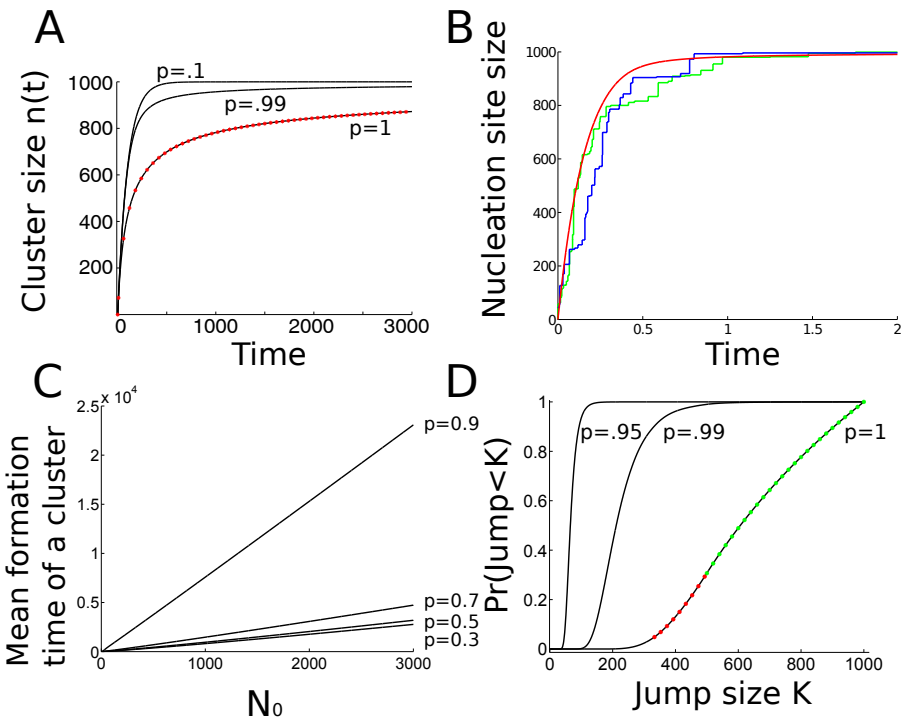

Figure 7: (A) Kinetics of the cluster growth in the mean field approximation. Various kinetics profile (solution of eq. (86)) for $N_T = 1,000$, $N_0 = 1,000$, $\lambda = 10^{-2}$ and $p = 0.1, 0.99, 1$. (B) Comparison of the kinetics of formation with the deterministic and the stochastic models. Parameters are $N_0 = 1000$, $p = 0.96$, $\lambda = 0.001s^{-1}$ (C) Mean time to form a function as a function of N_0 for various values of p, with $\lambda = 0.001s^{-1}$ and $N_T = 1000$. (D) Cumulative distribution of the maximal jump size $P_{N_0,K}$, for $N_0 = 1000$ and $p = 0.95, 0.97, 0.99, 1$. The probability $P_{N_0,K}$ for $p = 1$ is compared with the approximation eqs. (104) (green) and (105) (red).

membrane molecules. However, at low resolution, the motion is described by an effective stochastic equation [18, 20]

$$\dot{X} = a(X)dt + \sqrt{2}B(X)\dot{W}, \tag{107}$$

where $a(X)$ is the drift field and $B(X)$ the diffusion matrix. The effective diffusion tensor is given by $D(X) = \frac{1}{2}B(X)B^T(X)$ ($.^T$ denotes the transposition) [35, 34]. The observed effective diffusion tensor is not necessarily isotropic and can be state-dependent, whereas the friction coefficient γ in 106 remains constant and the microscopic diffusion coefficient (or tensor) may remain isotropic.

The drift field $a(\boldsymbol{x})$ in equation 107 represents a force that acts on the diffusing particle, regardless of the existence or not of a potential well [23]. In the case where $\boldsymbol{D}(\boldsymbol{x})$ is locally constant and the coarse-grained drift field $\boldsymbol{b}(\boldsymbol{x})$ is a gradient of a potential

$$a(\boldsymbol{x}) = -\nabla U(\boldsymbol{x}), \tag{108}$$

then the density of particles represents locally the Boltzmann density $e^{-U(\boldsymbol{x})/D}$ [23]. The force field can form potential wells, generically approximated locally as a paraboloid with an elliptic base. It remains a difficult question to extract the axis, the center and the boundary of the elliptic base of the well. Once they are known, within the analytical representation $U(\boldsymbol{x}) = A\left((\frac{x}{r_x})^2 + (\frac{y}{r_y})^2\right) + O(x,y)^2$, the constants A, r_x, r_y are three parameters to be determined.

GAG proteins show free and confined motions. The density of particles is quite heterogeneous, with many small dense regions and a few very dense regions (Fig. 8). The stochastic analysis and diffusion map (Fig. 8) reveals a mean diffusion coefficient is $D = 0.7\mu m^2/s$ almost uniform. Several potential wells could be detected with an elliptical base with radius 170 nm - 200 nm. One with depth $A = 0.78\mu m^2/s$ with a score of 0.20, confirming that these wells are robust [18]. The energy of the potential well is in the range 1.7-4 kT.

Interestingly, the wells evolve in time (Fig. 9) and can disappear rapidly (in less than 5 minutes) and the energy decreases gradually in time. This analysis used a moving window, which smooths out fluctuations. To observe the evolution of the trajectories in a small region in the proximity of the potential well, we plotted windows of 180 s of recording (Fig. 9). For each panel, the trajectories were recorded in the time interval $(t, t + 180s)$. The next panel represents trajectories taken 10 seconds afterwards, in the time interval $(t+10s, t+190s)$. To represent the evolution of trajectories through

time, in each window, the trajectories are colored during the first seconds in blue, and trajectories near 180 s in red. The most recent trajectories are overlaid on the first trajectories.

In the first seconds, trajectories are appeared unorganized (fig. 9, top row). Confined trajectories appear in only 10 seconds at 1750 s(Panel 1570–1750 s). This confinement lasts for 140 s: after time 1890 s (Panel 1710–1890 s), the new trajectories (red) that pass over the former confinement region are diffusive and not attracted to any point. To conclude, the potential well lasted for 140 s between 1750 s and 180 s. Moreover the confinement region is expanding through time. The radius of the potential well changed from approximately 200 nm at the beginning at 1750 s (Panel 1570–1750 s) to a radius of 250 nm at 1890 s. To measure the changes in energy of the well through time, the energy of the well in each window of 180 s is shown in fig. 9 lower panel. During the time interval (1750–1890 s), which corresponds to the period of confinement, the proximity of the measured drift map with a parabolic expression is in very good agreement in the confinement in the time period (1750–1890 s). This agreement confirms the presence of interaction forces acting on the Gag proteins. Finally, the present analysis confirms that aggregate formation occurs in geometrical confined structures that are transient in time.

5.7 Conclusion

We presented here several analytical formula based on aggregation-fragmentation with a finite of particles. These formulas can be used to extract parameters such as rate constants from experimental data. The general framework is also used to derive the extreme statistics about the time formation of a cluster or the time two particles spend in the same cluster.

We also discussed here two important applications about telomere clustering in yeast [17, 19] and capsid formation [22]. The geometrical organization of a cluster formation from small aggregates remains difficult to account for into modeling. Future directions should be concerned with accounting for the random geometry of aggregates and their insertion in a cluster. In the last subsection, we reviewed experimental evidences that capsid assembly might use the membrane local curvature, but the exact mechanism remain opens.

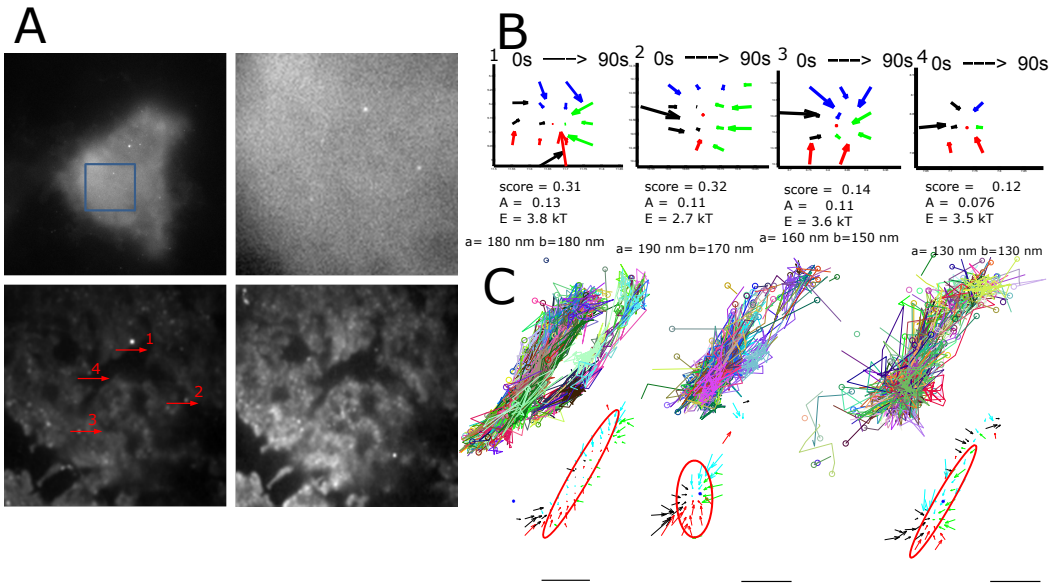

Figure 8: **A.** Area of aggregation (right) correspond to potential wells, characterized by converging arrows. **B.** 4 examples are shown (Left) with associated parameters of the ellipses: long a and short b abscise.**C.** three other potential wells represented with the high density of GAG trajectories, scale Bar 500nm (data given by the courtesy of S. Manley).

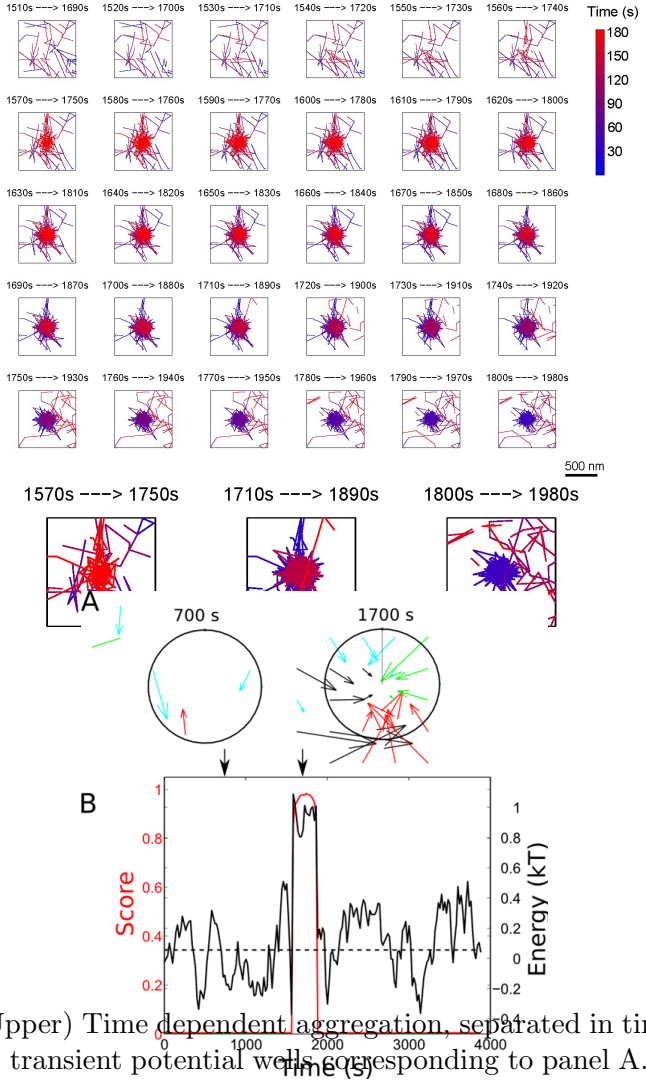

Figure 9: (Upper) Time dependent aggregation, separated in time windows. Lower **A-B.** transient potential wells corresponding to panel A. For the last panel, some GAG trajectories are not attracted by potential wells (data given by the courtesy of S. Manley).

Acknowledgements

This research was supported by a Marie-Curie grant. We thank S. Manley for discussions and sharing with us the GAG super-resolution data. We also thank the hospitality of the Newton Institute in Cambridge during the year 2016.

References

[1] Abramowitz, M. and Stegun, I.A. (1992). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* Reprint of the 1972 edition. Dover, New York.

[2] Aldous, D. J. (1999). Deterministic and stochastic models for coalescence (aggregation, coagulation): review of the mean-field theory for probabilists. *Bernoulli* **5** 3–48.

[3] Andrews, G.E. (1976). *The Theory of Partitions. Encyclopedia of Mathematics and its Applications, Vol. 2.* Addison-Wesley, Reading, MA.

[4] Ball, J. M. and Carr, J. (1990). The discrete coagulation-fragmentation equations : existence, uniqueness and density conservation. *J. Stat. Phys.* **61** 203–234.

[5] Becker, R. and Döring, W. (1935). Kinetische Behandlung der Keimbildung in übersättigten Dämpfen. *Ann Phys* **24** 719–752.

[6] Chandrasekar, S. (1943). Stochastic problems in physics and astrophysics. *Rev. Mod. Phys.* **15** 1–89.

[7] Collet, J.F. (2004). Some modelling issues in the theory of fragmentation-coagulation systems. *Commun. Math. Sci.* **1** 35–54.

[8] Bystricky, K., Heun, P., Gehlen, L., Langowski, J., and Gasser, S. M. (2004). Long-range compaction and flexibility of interphase chromatin in budding yeast analyzed by high-resolution imaging techniques. Proceedings of the National Academy of Sciences of the United States of America, 101(47), 16495-16500.

[9] Carlsson, T., Ekholm, T., and Elvingson, C. (2010). Algorithm for generating a Brownian motion on a sphere. Journal of physics A: Mathematical and theoretical, 43(50), 505001.

[10] Doering, C.R. and Ben-Avraham, D. (1988). Interparticle distribution functions and rate equations for diffusion-limited reactions. *Phys. Rev. A* **38** 3035.

[11] D'Orsogna, M. R., Lakatos, G. and Chou, T. (2012). Stochastic self-assembly of incommensurate clusters. *J. Chem. Phys.* **136** 084110.

[12] D'Orsogna, M. R., Lei, Q. and Chou, T. (2015). First assembly times and equilibration in stochastic coagulation-fragmentation. *J. Chem. Phys.* **143** 014112.

[13] Durrett, R., Granovsky, B.L. and Gueron, S. (1999). The equilibrium behavior of reversible coagulation-fragmentation processes. *J. Theoret. Probab.* **12** 447–474.

[14] Gueron, S. (1998). The steady-state distributions of coagulation-fragmentation processes. *J. Math. Biol.* **37** 1–27.

[15] Gunzenhäuser J, Wyss R, Manley S, 2014. A quantitative approach to evaluate the impact of fluorescent labeling on membrane-bound HIV-Gag assembly by titration of unlabeled proteins. *PLoS One.* 9(12): e115095.

[16] Holcman, D., Schuss, Z. (2015). Stochastic narrow escape in molecular and cellular biology: analysis and applications. Springer.

[17] Hoze, N. and Holcman, D. (2012). Coagulation–fragmentation for a finite number of particles and application to telomere clustering in the yeast nucleus. *Phys. Lett. A* **376** 845–849.

[18] Hoze N, Nair D, Hosy E, Sieben C, Manley S, Herrmann A, Sibarita JB, Choquet D, Holcman D, 2012. Heterogeneity of receptor trafficking and molecular interactions revealed by superresolution analysis of live cell imaging, *Proc. Natl. Acad. Sci. USA* 109: 17052–17057.

[19] Hoze, N., Ruault, M., Amoruso, C., Taddei, A. and Holcman, D. (2013). Spatial telomere organization and clustering in yeast Saccharomyces cerevisiae nucleus is generated by a random dynamics of aggregation–dissociation. *MBoC* **24** 1791–1800.

[20] Hoze N, Holcman D, 2014. Residence times of receptors in dendritic spines analyzed by stochastic simulations in empirical domains. *Biophys. J.* 107:3008-17.

[21] Hoze, N. and Holcman, D. (2014). Modeling capsid kinetics assembly from the steady state distribution of multi-sizes aggregates. *Phys. Lett. A* **378** 531–534.

[22] Hoze, N. and Holcman, D. (2015). Kinetics of aggregation with a finite number of particles and application to viral capsid assembly. *J. Math. Biol.* **70** 1685–1705.

[23] Holcman D, Hoze N, Schuss Z, 2015. Analysis and interpretation of superresolution single-particle trajectories. *Biophys. J.*, 109:1761-1771.

[24] Hoze, N. and Holcman, D. (2016). Stochastic coagulation-fragmentation processes with a finite number of particles (submitted).

[25] Jacquot, S. (2009). A historical law of large numbers for the Marcus-Lushnikov process. *Electron. J. Probab.* **15** 605–635.

[26] Jones, W.B. and Thron, W. J. (1980). *Continued Fractions: Theory and Applications.* Addison-Wesley, Reading, MA. pp. 198-214.

[27] Kelly, F.P. (1979). *Reversibility and Stochastic Networks.* Wiley Series in Probability and Mathematical Statistics, Chichester.

[28] Krapivsky, P., Redner, S. and Ben-Naim, E. (2010). *A Kinetic View of Statistical Physics.* Cambridge University Press, Cambridge.

[29] Liggett, T. (1985). *Interacting Particle Systems.* Springer, New York.

[30] Lushnikov, A. A. (1978). Coagulation in finite systems. *J. Colloid Interface Sci.* **65** 276–285.

[31] Marcus, A. (1968). Stochastic coalescence. *Technometrics* **10** 133–143.

[32] Meyer Jr., C.D. (1975). The role of the group generalized inverse in the theory of finite Markov chains. *SIAM Rev.* **17** 443–464.

[33] Rotstein, H. G. (2015). Cluster-size dynamics: A phenomenological model for the interaction between coagulation and fragmentation processes. *J. Chem. Phys.* **142** 224101.

[34] Schuss Z, 2010. *Theory and Applications of Stochastic Processes: An Analytical Approach.* Applied Mathematical Sciences vol.170, Springer NY.

[35] Schuss Z, 2011. *Nonlinear Filtering and Optimal Phase Tracking.* Applied Mathematical Sciences vol. 180, Springer NY.

[36] von Smoluchowski, M. (1916) Drei Vorträge über Diffusion Brownsche Molekularbewegung und Koagulation von Kolloidteichen. *Physik Z* **17** 557-571.

[37] Szegö, G. (1975). *Orthogonal polynomials*, 4th edition, Amer. Math. Soc. Colloq. Publ. Providence, RI. p. 198.

[38] Thompson, C. J. (1988). *Classical Equilibrium Statistical Mechanics.* Oxford University Press, Oxford.

[39] Thomson, B. R. (1989). Exact solution for a steady-state aggregation model in one dimension. *J. Phys. A* **22** 879–886.

[40] Wattis, J. A. (2006). An introduction to mathematical models of coagulation–fragmentation processes: a discrete deterministic mean-field approach. *Phys. D* **222** 1–20.

[41] Yvinec, R., D'Orsogna, M.R. and Chou, T. (2012). First passage times in homogeneous nucleation and self-assembly. *J. Chem. Phys.* **137** 244107

[42] Zlotnick, A. (2005). Theoretical aspects of virus capsid assembly. *J. Mol. Recognit.* **18** 479–490.

[43] Schuss, Z. (2010). Diffusion and Stochastic Processes: an Analytical Approach. (Springer, New York).

[44] Matkowsky, B.J., Z. Schuss, C. Knessl, C. Tier, and M. Mangel. (1984). Asymptotic solution of the Kramers-Moyal equation and first passage times for Markov jump processes. Phys. Rev. A. 29, 3359.

[45] Gillespie, D.T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J. Comput. Phys. 22, 403.

[46] Doi, M., Edwards, S. F. (1988). The theory of polymer dynamics (Vol. 73). Oxford university press.

[47] Therizols P, Duong T, Dujon B, Zimmer C, Fabre E, (2010). Chromosome arm length and nuclear constraints determine the dynamic relationship of yeast subtelomeres. Proc Natl Acad Sci U S A. 107 :2025–30.

[48] Schober H, et al. Controlled exchange of chromosomal arms reveals principles driving telomere interactions in yeast. Genome Res. 2008;18:261–271