Microsatellites distribution within Smoluchowski equation approach

Maksym Serbyn
(Dated: May 7, 2010)

We present a different approach to the existing models [1–3] for equilibrium lenght distribution of microsatellites. We show that continuous form of master equation coincides with Smoluchowski equation with diffusion and fragmentation processes. Solving this equation we obtain analytic form of equilibrium distribution. This approach allows us to distinguish essential features of initial stochastic models from inessential ones. We also discuss possibility of inclusion of merging processes and its consequences.

ʊ

I. INTRODUCTION

Microsatellites are repeats of short patterns occurring in DNA. They are very important as genetic markers, moreover they are believed to be responsible for certain diseases in humans. One address many different aspects of microsatellites evolution, see e.g. [4]. Here we concentrate solely on length distribution of microsatellites.

Even having such a restriction one can address many different aspects. Namely, we can study *equilibrium* length distribution of microsatellites. The simplest model [5] includes only diffusion in the space of length and has no equilibrium distribution. Kruglyak et al. in [1] and [2] generalized this model including point mutation processes. Their model is Markov chain evolution. Further generalization of Kruglyak's model has been made in [3]. Authors not only extend Kruglyak's model, but also introduce genetic distances. After this they study *time evolution* of genetic distance and use it as a test for various models. In other words, they consider time dependence of various correlators. Computing these correlators from experimental data, one can get estimate for the time and compare it with experimentally established value.

One can observe that in all models of microsatellite evolution, dynamics is governed by processes of diffusion and fragmentation. The same processes along with merging are very common in many physical systems. For example, in [6] authors study Smoluchowski equation for the distribution function $N(x,t)$ for crystal sizes. They include merging and coagulation terms in master equations and find equilibrium distribution. For specific forms of kernels authors find scale invariant solutions. Moreover, equilibrium solution found by authors is in good agreement with experimental data for different systems [7].

In what follows we will apply approach of [6] to the Kruglyak's model of microsatellite evolution. First, in SEC. II we will review existing models for microsatellite evolution and discuss their predictions. Then, in SEC. III we will apply methods of [6] to the Kruglyak's model of microsatellites distribution. In SEC. IV we will discuss possible merging kernels and their consequences. Finally, we will repeat our main results and conclusions in the last section.

II. REVIEW OF EXISTING MODELS

Length of microsatellites can change during the process of replication [4]. There are two possible processes: *polymerase slippage*, that can be modeled as the diffusion in the space of length and much less probable *point mutations*. Note, that in the process of polymerase slippage the length or microsatellite changes by introducing/removing one nucleotide. While in the point mutation process, certain nucleotide changes, thus breaking microsatellite into two.

A. Model of [1] and [2]

In [1, 2] Kruglyak et al. introduce Markov chain model with following dynamics. The length of microsatellite $N_t = \ell$ in the next step may become:

- $\ell \to \ell + k$ at rate $r_{\ell,k}$ (polymerase slippage)
- $\ell \to j$ where $1 \le j < \ell$ at rate a (point mutations)
- $\ell = 1 \to 2$ at rate c (start of a new repeat).

$r_{\ell,k}$ is taken to be proportional to the length, i.e. in the form $r_{\ell,k} = (\ell - 1)b_k$ where b_k is non-zero only when $|k| < K$ where K is some fixed number.

Having formulated explicit model, Kruglyak et all have surprisingly almost no *explicit analytic* results for it. Namely, they give mathematical proof of the existence of stationary distribution and give two *estimates* (from above) for average length of microsatellites. In addition, they find exact solution for model where $r_{\ell,-1} = (\ell-1)d$ for $\ell \ge 2$ and $r_{\ell,1} = \ell e$ for all $\ell \ge 1$. The resulting distribution is geometric. Using numerical simulations, they obtain distribution for model with $b_1 = b_{-1} \ne 0$ and compare it with the data. They compare only one parameter, mutation rate per locus, obtaining the same order of magnitude as in experiment (but difference is still factor of 2-4).

B. Model of [3]

In [3] Calabrese et al. consider a variety of different models. Namely, authors consider SMM model (step-

wise mutation) introduced in [5], PS/PM model (proportional slippage/point mutations) by Kruglyak and it's slight generalization — PCR model that keeps track of the whole ensemble of microsatellites rather than certain one. The tool that authors use to compare all these models are *genetic distances*. Authors compare model predictions for *time dependence of genetic distances* with the data. As a summary they have following models and results:

- SMM model — diffusion, microsattelites change by ± 1 unit at a rate β independent on their length. Authors find exact dependence of mean square genetic distance on time (it coincides with the results for random walk).
- PS/PM model — model of Kruglyak et al. [1], with the minimal length of microsatellite equal to κ rather than 1. The only nonzero b are $b_{\pm 1} = b$ and $r_{\ell,\pm 1} = (\ell - \kappa)b$. Correspondingly, birth process is $\kappa \to \kappa + 1$ at rate c.
- PCR model — designed to describe PCR results more appropriately. It deals with the evolution of vector $(X_t^1, \dots X_t^n)$, where each X^i is microsatellite and evolves according to the Kruglyak's model. The point mutations decrease number of components of this vector by one. Results for this model include expression for variance of the length of microsatellites.
- PS/0M model — model of Kruglyak et al for total length of microsatellites *without point mutations*. It is expected to be valid for small times. It occurs to be equivalent to so-called binary branching process Z_t of probability theory. Authors compute dependence of moments on time and probability distribution.

C. Summary

Despite the existence of a number of explicit models, there are only very few explicit results that are formulated only for simplest cases. Moreover, although computer simulations allow us to obtain distributions, they give little or no information about properties of these distributions as well as about sensitivity of final results to various modifications of initial model.

III. KRUGLYAKS MODEL: ANALYTIC CONSIDERATION

A. Master equation

We start with the simplest version of Kruglyak's model, described in SEC II A. Namely, we have 2 processes [8]:

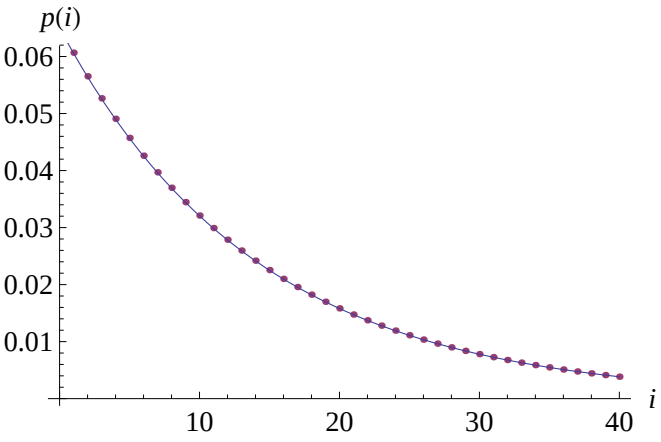

FIG. 1. Comparison of equilibrium distribution $p(i)$ obtained from continuous and discrete versions of master equation. Coefficients in Eq.(3) are chosen to be $a = 0.0001$ and $b = 0.02$.

$$l \to l \pm 1 \quad \text{at rate} \quad lb \tag{1}$$

$$l \to j \quad \text{at rate} \quad a. \tag{2}$$

For such process one can readily write down the Master equation

$$\frac{d}{dt}p(l) = b(l+1)p(l+1) + b(l-1)p(l-1) - 2blp(l) + \sum_{k=l+1} 2ap(k) - a(l-1)p(l). \tag{3}$$

Here first three terms describe diffusion process while last two terms correspond to fragmentation processes. The fragmentation term does not conserve total probability, however, it conserves total length. This imposes certain restrictions on the diffusion coefficients, if we want the diffusion terms to conserve total length. In [6] authors use constant diffusion coefficient. The only possible generalization of it is *symmetrical* length dependent diffusion coefficient. I.e. we require rates of processes $l \to l \pm 1$ to be equal, although they may depend on l.

B. Continuous version

Discrete master equations are very difficult to solve explicitly. Thus, in order to solve Eq. (3) we go to continuous version. This is done by introducing lattice spacing Δ and new variable $x = \Delta l$. Then, denoting

$$b\Delta = D, \qquad \frac{a}{\Delta} = f, \tag{4}$$

one obtains

$$\frac{d}{dt}p(x) = D\frac{d^2}{dx^2}xp(x) - fxp(x) + 2f\int_x^\infty dy\, p(y). \tag{5}$$

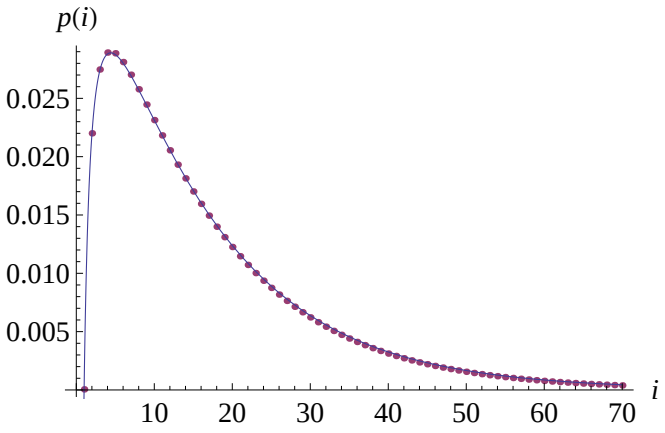

FIG. 2. Comparison of equilibrium distribution $p(i)$ obtained from continuous and discrete versions of master equation. Coefficients in Eq.(3) are chosen to be $a = 0.0001$ and $b = 0.02$. The boundary condition at $i = 1$ is absorbing, $p(1) = 0$.

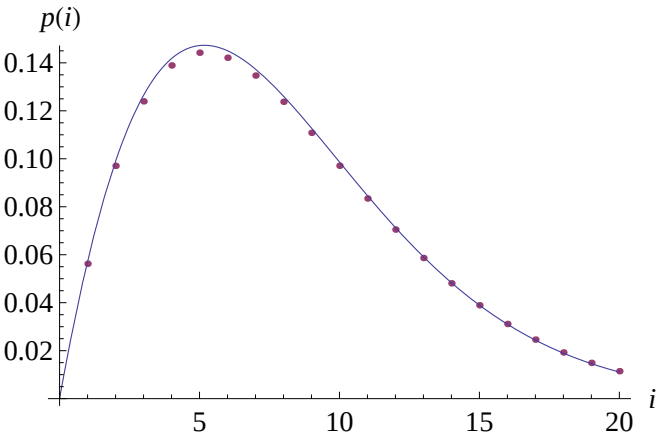

FIG. 3. $p(i)$ for constant diffusion coefficient, $a = 0.0001$, $b = 0.02$.

This equation is almost identical to the one considered in [6] with the only difference that we have an extra x in the term with the second derivative. Origin of factor x is clear: it comes from the length l that is present in rate (1). Differentiating Eq. (5) we obtain differential equation of third order that can easily be solved. The solution can be written as

$$p(y) = C_1 e^{-y} + C_2 e^{y} + C_3 \left[e^{y}\mathrm{Ei}(-y) - e^{-y}\mathrm{Ei}(y) + \frac{2}{y} \right], \tag{6}$$

where y is rescaled variable,

$$y = x\sqrt{\frac{f}{D}} = \frac{x}{\Delta}\sqrt{\frac{a}{b}}, \tag{7}$$

and $\mathrm{Ei}(x)$ is exponential integral function, defined as

$$\mathrm{Ei}(x) = -\mathcal{P}\int_{-x}^{\infty} dt\, \frac{e^{-t}}{t}. \tag{8}$$

In order fix C_1, C_2, and C_3 we need to specify boundary conditions. First is obvious: it is vanishing $p(x)$ when $x \to \infty$, which sets $C_2 = 0$. Second boundary condition is less trivial, since our equation is singular at zero. In case of reflecting boundary, $p(x)$ is monotonic, and one can infer that C_3 vanish, thus we are left with simple exponential solution (see FIG. 1):

$$p(y) = C_1 e^{-y}. \tag{9}$$

However, if we have absorbing boundary and $p(\Delta) = 0$, we can find C_1 and C_3 from normalization and regularized boundary condition $p(\Delta) = 0$. Thus we have more interesting solution, with slowly decaying *power-law tail* (see FIG. 2):

$$p(y) = C_1 \left(e^{1-y} - \frac{\tilde{C}_2}{2}\left[e^{y}\mathrm{Ei}(-y) - e^{-y}\mathrm{Ei}(y) + \frac{2}{y} \right] \right) \propto \frac{1}{y^3} + O\left(\frac{1}{y^5}\right), \tag{10}$$

where C_1 is again determined from normalization and $\tilde{C}_2 \approx 1$ from the condition $p(y = 1) = 0$.

From FIG. 1-2 one can see that analytical solutions for continuous version of master equation are in good agreement with numerical simulations. In order to see how this machinery works for different equation, on FIG. 3 we also plot the numerical and analytical solutions to the master equation with constant diffusion coefficient:

$$\frac{d}{dt}p(l) = bp(l+1) + bp(l-1) - 2bp(l) + \sum_{k=l+1} 2ap(k) - a(l-1)p(l). \tag{11}$$

Resulting continuous equation are identical to the one, used in [6]. Its solution and asymptotic at $y \to \infty$ has the form:

$$p(y) = C_1 y^{3/2} K_{\frac{1}{3}}\left(\frac{2y^{3/2}}{3}\right) \propto y^{3/4} e^{-\frac{2y^{3/2}}{3}}. \tag{12}$$

To summarize, in contrary to exact solutions found in [1–3], continuous version of master equation allows to understand which aspects of our model have essential impact on our equilibrium distribution and deduce its asymptotic behavior. Namely, one easily can see that fact that we used lb instead of $(l-1)b$ as a diffusion rate have almost no impact. Whenever change of dependence of diffusion rate from being proportional to l to constant, changes asymptotic behavior and form of distribution function (compare Eqs. (10) and (12)).

IV. MERGING KERNELS

If we consider merging due to the point mutation processes, it's rate will be negligible, since it is proportional to the probability to have exactly reverse mutation. Still one can try to consider the influence of merging kernels on the equilibrium distribution. This has been done in [6, 9], here we will apply their results to our equation. The general continuous master equation with merging kernel has a form:

$$\partial_t p(x) = D\partial_x^2 x p(x) + [\partial_t p(x)]_{\text{frag}} + [\partial_t p(x)]_{\text{merg}}; \quad (13)$$

$$[\partial_t p(x)]_{\text{frag}} = -p(x)\int_0^x dx'\, F(x-x',x') + 2\int_0^\infty dx'\, F(x,x')p(x+x'); \quad (14)$$

$$[\partial_t p(x)]_{\text{merg}} = -p(x)\int_0^\infty dx'\, K(x-x',x')p(x') + \frac{1}{2}\int_0^x dx'\, K(x-x',x')p(x-x')p(x'). \quad (15)$$

One see that with constant fragmentation kernel $F(x,x') = f$ and zero merging kernel we reproduce Eq. (5). One can consider merging kernel of degree λ, i.e. such that

$$K(ax,ax') = a^\lambda K(x,x'), \quad (16)$$

along with present fragmentation kernel of degree $k = 0$. Then, when $\lambda < 2$ one will have stationary solution, while for $\lambda > 2$ long microsatellites will be dominating and thus continuous equation will be not applicable. At the critical value of $\lambda_c = 2$ stationary scale invariant solution can exist (see details in [6]). Comparing this with experimental data [1, 3], we see that possible merging kernels could have degree $\lambda < 2$. However, even with such a kernel, mean field results can not be considered as reliable. According to the [9], upper critical dimensionality of the systems is $d_c = 2$. For $d < d_c$ fluctuations of spatial density make the behavior of the system different from mean-field. Moreover, based on the model of microsatellites as a repeat sequences in DNA, we see that merging kernel will destroy Markovian property of the model. Therefore, possible consideration of merging kernel is possible only within numerical simulations.

V. CONCLUSIONS

To conclude, we analyzed existing models for the stationary distribution of microsatellite length. Starting from master equation we obtained (continuous) Smoluchowski-type of equation for the model presented in [1, 2]. We found its analytical solution for different types of boundary conditions and checked validity of our result, by comparing it to the numerical solution of master equation. Finally, we discussed possibility of inclusion of merging kernel into Smoluchowski equation.

[1] R. Durrett and S. Kruglyak, J. Appl. Probab. **36**, 621 (1999).

[2] S. Kruglyak, R. T. Durrett, M. D. Schug, and C. F. Aquadro, Proc Natl Acad Sci U S A **95**, 1077410778 (1998).

[3] P. P. Calabrese, R. T. Durrett, and C. F. Aquadro, Genetics **159**, 839852 (Oct 2001), http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1461831/.

[4] C. Schlötterer, Chromosoma **109**, 365 (Sep 2000).

[5] M. Kimura and T. Ohta, PNAS **75**, 2868.

[6] J. Ferkinghoff-Borg, M. H. Jensen, J. Mathiesen, and P. Olesen, Physica D: Nonlinear Phenomena **222**, 88 (2006).

[7] J. Ferkinghoff-Borg, M. H. Jensen, J. Mathiesen, P. Olesen, and K. Sneppen, Phys. Rev. Lett. **91**, 266103 (Dec 2003).

[8] Note, that we write lb instead of $(l-1)b$ in order to avoid introduction of constant c that corresponds to the birth rate of microsatellites.

[9] P. Meakin and M. H. Ernst, Phys. Rev. Lett. **60**, 2503 (Jun 1988).