Boulesteix, Hösel, Liebscher:

Stochastic modeling for the COMET-assay

Stochastic modeling for the COMET-assay

A.-L. Boulesteix[*] V. Hösel[†] V. Liebscher[‡]

October 6, 2003

Abstract

We present a stochastic model for single cell gel electrophoresis (COMET-assay) data. Essential is the use of point process structures, renewal theory and reduction to intensity histograms for further data analysis.

1 Introduction

Single cell gel electrophoresis or "COMET-assay" is a very efficient method to examine DNA damage and repair with many applications, for example in cancer research. A non-damaged DNA molecule is a long linear chain of desoxyribonucleic acids. When a cell is irradiated several strand breaks in the DNA may occur. The aim of the study is to detect to which amount a broken DNA molecule can be recombined by the organism. Non efficient repair may indicate genetically determined malfunctions in the recombination and replication mechanisms of the DNA. At present, the COMET assay is the only technique to monitor DNA damage and repair at the level of single cells.

The standard way to analyze COMET data is to compute characterizing geometric properties of the comet, e.g. the tail moment [3] or the comet moment [19]. Some of these parameters show only little variability across experiments [12]. However, all these

[*] Ludwig-Maximilian-Universität Munich, Dept. of Statistics, Ludwigstr.33, D–80539 Munich, Germany, email:socher@stat.uni-muenchen.de

[†] Munich University of Technology, Centre of Mathematics, D–80539 Munich, Germany, email:hoesel@t-online.de

[‡] GSF — National Research Centre for Environment and Health, Institute of Biomathematics and Biometry, Ingolstdter Landstr.1, D–85758 Neuherberg, Germany, email:liebscher@gsf.de

parameters are sensible to small changes in the recorded COMET image. The used image processing method plays a role and the errors in detecting faster fragments. Such fragments appear usually darker and are thus not well separable from the background of the image. Further, the images contain much more information than could be coded in one or a few parameters. If one had a comprehensive model for the whole data and some robust methods to extract relevant information, one could make better use of the recorded images. In the present work, we show that such modelling and also robust classification of the comets is possible. On the other hand, extracting more information from the images on the basis of the model leads to more subtle empirical deconvolution problems. This will be addressed in a forthcoming paper [7].

This work is structured as follows. Section 2 is a short introduction to the COMET-assay and to our approach for its modelling. In section 3 we will describe the problem as a marked point process. In the subsequent sections we derive stochastic models for the various stages of the experiment like the distribution of fragment masses after radiation and the mass dependent migration distance of a single DNA fragment. In section 7, we discuss the combined model by means of simulation and parameter estimation.

2 The COMET-assay and its modelling

A large amount of articles describe the technical details of the COMET-assay, for instance [3, 6, 19]. An up to date source of information is the Web site [17] and an extensive review of the COMET-assay can be found in [13]. Here, we only present a short overview of the method with emphasis on few features which are important for our mathematical modelling.

The cells to analyze are attached to an agarose gel and placed in an electric field, after suitable treatment and in particular conditions. Since DNA is polar, DNA molecules tend to migrate. Big DNA molecules (i.e. non-damaged or repaired DNA molecules) show no observable migration, whereas small DNA molecules (i.e. damaged DNA molecules) migrate quickly off the center of the cell. These small fragments constitute the tail of the comet like electrophoresis image. Hence the name COMET-assay, see Figure 1. It is quite difficult to explain why small molecules migrate faster than big ones, but one of the main explanations is that big molecules are more sensitive to hurdles (gel fibers) during the migration. Till now there are diverse opinions among biologists about the underlying mechanisms. Anyway, at the end of the electrophoresis, it is possible to see whether a cell is ’quite damaged’ or ’quite non-damaged’, by analyzing the shape of the comet: a damaged cell has a long and/or dense tail, whereas a non-damaged cell merely looks like a homogeneous disk.

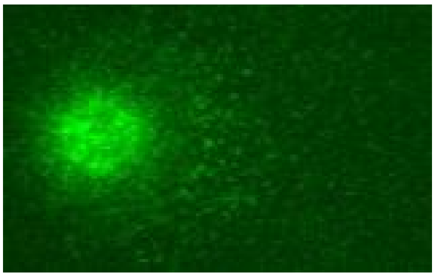

Figure 1: A comet from an irradiated cell

Our aim in the present paper is to establish a reasonable stochastic model describing the data which can serve as a basis for future statistical inference. This strategy is in contrast to the standard approach, which uses only a few geometric features. We emphasize again that the image data contain *much information* not represented in the single geometric parameters. We are able to retrieve this information only if we can model the physical processes of the experiment with sufficient accuracy. The final goal is to estimate the distribution of lengths of DNA molecules (or equivalently their distribution of mass) in damaged and repaired cells, in order to get more information on the repair mechanism and its efficacy. Keeping track of the approximations and assumptions in the modelling process will help to implement methods which are robust under changes of model parameters and slight violations of model assumptions.

In terms of the data, the distribution of molecule lengths we want to estimate is best associated to the distribution of displacements of single DNA molecules. This demands some further knowledge about the relation between length and speed. In the literature, biologists propose theoretical models (for instance in [27]) for this relation and give experimental results (for instance in [23]) obtained in various conditions. These studies are especially designed for usual gel electrophoresis, where the lengths of the DNA fragments is ≈ 500 bp. This is much smaller than the fragment lengths considered in COMET experiments. In section 3 we propose a global model to describe the DNA migration and finally get a formula agreeing with some of the experimental results. Our model takes into account a great part of the physical features cited in the literature and is quite consistent with the empirical formulae already known.

Our model for the available data is guided by the experiment: First we describe the placement of the DNA fragments before radiation and after radiation. To model the effect of the gel electrophoresis, we then give a mathematical description of the migration of DNA molecules through an agarose gel.

3 Marked Point Processes as Description

We consider a single cell containing N DNA fragments, where N is a (random) number depending on the number of DNA breaks. Each of the N fragments is represented by a tuple $(\mathbf{X}_i, m_i)$, $i \in \{1, \ldots, N\}$, where $\mathbf{X}_i$ is a three dimensional vector representing the initial location of fragment i and m_i is its mass. $\Xi = \{(\mathbf{X}_i, m_i) : i \in \{1, \ldots, N\}\}$ corresponds to the observed fragments, approximating the location of a fragment by a point, but carrying its mass into the calculations via m_i. Note, that we can not differentiate between break experiments resulting in fractions of the same size. So, the set Ξ which is a simple finite marked *point process* [11] is a natural description of the fragments. Let $\mathbf{D}_i$ be the three dimensional vector of displacement of the i-th fragment and $\mathbf{X}'_i = \mathbf{X}_i + \mathbf{D}_i$, which is thus the three dimensional vector of end location of fragment i. $\mathbf{X}$, $\mathbf{D}$ and $\mathbf{X}'$ are depicted in figure 2 for one point.

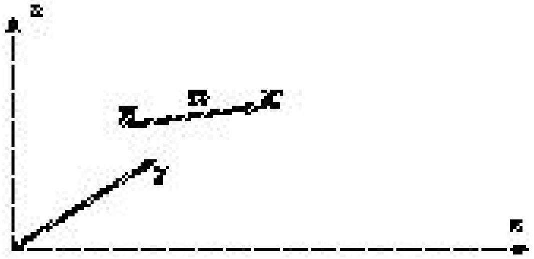

Figure 2: Coordinate system. The represented point is at $\mathbf{X}$ at the beginning, its displacement is $\mathbf{D}$ and its end location is $\mathbf{X}'$.

4 The Length Distribution — Poisson Approximation

Our goal is to determine the distribution of fraction lengths. It remains a very complex issue because we do not know much about the mechanisms of breakage and repair.

However, with a few simple assumptions, one can regard the distribution of lengths in a damaged cell as exponential. This model is commonly called ’Random Breakage Model’ (RBM) and described in [21]. Let us briefly discuss its underlying assumptions.

1. Breaks occur completely at random, i.e. the radiation causes only single break events which do not influence each other (this hypothesis is supported by the low energy of γ-radiation) .

2. Breaks occur homogeneously, i.e. no part of the the DNA strand has a higher or lower risk for break events. This hypothesis is questionable, since the DNA in the cell nucleus has spatially a very complicated crystalline structure and parts of DNA deeper inside this structure may be exposed to less radiation.

In our setting it is also sensible to make the following additional assumptions

- Breaks are rare compared to the number of unbroken sites.

  From literature, we know that depending on experimental conditions 1 Gy radiation intensity causes on average one single strand break (ssb) every several ten thousand base pairs (bp). For example, [24] gives a value of $5.98*10^{-8}Gy^{-1}Da^{-1}$, corresponding approximately to one ssd for every 25000 bp (with 660 Ga as average molecular weight of a bp). In our case, a 3.5 Gy γ-source has been used leading to a rough estimate of one ssb for every 7000 bp.

  Further, the considered mouse chromosomes exceed by far 10 Mbp and thus we state that

- the total number of breaks is large.

The following lemma gives a hint how to find distributions that model large numbers of rare events.

Lemma 1 *Suppose* $(N_k)_{k\in\mathbb{N}}$ *are random variables geometrically distributed with survival probabilities* $(q_k)_{k\in\mathbb{N}}$, $\lim_{k\to\infty} -k\ln q_k = \lambda$. *Then*

$$\mathcal{L}(\frac{N_k}{k}) \xrightarrow[k\to\infty]{} \mathrm{Exp}_\lambda.$$

If $(\Xi_k)_{k\in\mathbb{N}}$ *are simple point processes on* $\mathbb{N}$ *such that* $P(\{n_1,\dots,n_l\}\subseteq\Xi_k) = (1-q_k)^l$, $\lim_{k\to\infty} -k\ln q_k = \lambda$ *then*

$$\mathcal{L}(\Xi_k/k) \xrightarrow[k\to\infty]{} \Pi_\lambda,$$

where Π_λ *is the stationary Poisson process on* $\mathbb{R}_+$ *with intensity* λ *and* $\Xi/r = \{x/r : x \in \Xi\}$.

Proof. By use of Laplace Transform, [11, Proposition 9.1.VII]. □

The lemma tells us, that for all sufficiently long pieces of DNA the number of breaks can be be regarded as Poisson distributed with parameter $\nu_L|I|$, where ν_L is some strictly positive real constant and $|I|$ is the length of the piece.

Under the second assumptions, the length resp. the mass between two breaks is exponentially distributed with parameter ν_L resp., say, ν_M . In the following, we will always consider only the mass m. Note, that it would be strictly equivalent to consider the length instead. Thus, we can assume that the density of mass of the DNA fragments has the form

$$f_M(m) = \nu_M \exp(-\nu_M m), \qquad (m \geq 0).$$

This model suggests that we only have to determine the constant ν_M in order to know the distribution of mass completely. Indeed, one can find in the literature tables recording the average number of breaks per thousand of DNA bases for specific experimental conditions (including the radiation intensity). These numbers could in principle be used to determine the parameter of the exponential distribution. However, these tables are highly dependent on experimental conditions, which unfortunately do not fit our case. Therefore, we need the COMET-assay to fix the parameter ν_M.

Things get more complicated, if we consider the repair mechanism, which controls the data for the "repair" group. We assume

1. that breaks are repaired independently,

2. the repair mechanism is homogeneous (it does not depend on the site of the chromosome where the break occurred) and

3. there is no difference for the cell to repair breaks between short or long fragments.

This means, we assume that breaks are deleted independently of each other. In the language of point processes the process of break points is *thinned*.

The following lemma is well-known.

Lemma 2 ([11, Example 8.2(a)]) *If* Z *is a Poisson process with intensity measure* μ *then the thinned configuration* Z_p*, where each point of* Z *is deleted with probability* $0 \leq p \leq 1$ *is Poisson distributed with intensity measure* $p\mu$.

Due to the above assumption 3, this carries over to the case of marked point processes.

In our application *p* describes the repair efficacy. Of course, an estimation of *p* would be highly dependent on the RBM and our assumptions on the repair mechanism. So, we should look for robust substitutes of *p*.

5 Migration of a Single Fragment — Renewal Processes and Diffusion Approximation

In this part we propose a model for the migration of DNA fragments in order to determine theoretically the conditional distribution of displacement given the mass of a fragment. In the literature, various qualitative models have already been proposed, for instance DNA-fragments as small balls moving in a thin net of points ([4], [9], [28]) or as long 'snakes' creeping between big obstacles ([1], [23], [27], [28]). Most of these models are mainly qualitative and tailored for specific experimental conditions. Our model is an adaptation of the Ogston theory [9]. We aim to make a model simple enough to allow mathematical treatment, while taking into account as many features as possible.

Let us consider each cell separately. Since in our case the agarose gel concentration is very low, the gel can be assumed as a net of randomly distributed points, like in the Ogston model. DNA fragments are considered as solid round balls because we assume them to be rolled and not stretched. The modelling can easily be generalized if DNA fragments are assumed to be ellipsoids, as it was suggested in [4]. We admit, that this assumption on the geometry of the fragments is quite bold and this issue is still very controversial. But, it allows us simple modelling, which would be not possible without restrictive assumptions.

The whole cell is assumed to be a flat cylinder (See figure 3): each fragment can move in a three-dimensional space, but in fact we will not pay attention to the displacements along the vertical axis *z* because they are negligible in comparison with the displacements induced by the electric field, which is parallel to the *x*-axis. For the same reason we also neglect the displacements of the fragments in the *y*-direction. The displacements in the *y*-direction and *z*-direction are quite complicated to understand and to model. To sum up, they can be assumed approximately as complex diffusion movements. The displacement in the *x*-direction depends on the mass (or length) of the fragment in a way that will be specified later. In our model we will only consider the displacements in this direction.

This means, that we look for the projections on the x-axis of the vectors $\mathbf{X}$, $\mathbf{D}$ and $\mathbf{X}'$, which will be simply denoted as X, D and X'.

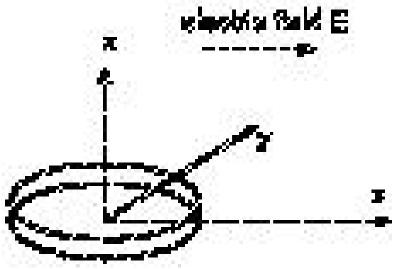

Figure 3: A whole cell in the coordinate system

Let us consider a fixed fragment i. When the electric field is applied, fragment i begins to migrate freely with constant and mass-independent speed (denoted v_0) in the x-direction during a period T_{i1} till it collides with an hurdle (gel fiber). Then it needs some time (S_{i1}) to bypass it, using the shortest path (see figure 4). Then it can migrate freely again during T_{i2} till it meets the next hurdle, etc. Thus the migration consists of a succession of periods $T_{i1}, S_{i1}, T_{i2}, S_{i2}, \ldots, T_{ik}, S_{ik}, \ldots$ The electric field is applied at time $t = 0$ and time t_0 corresponds to the end of the experiment, the time at which we observe the location of the fragment. The T_{ik} and S_{ik} can be seen as realizations of random variables T_k and S_k. In the following, we model the distribution of the $(T_k)_{k=1}^{\infty}$ and $(S_k)_{k=1}^{\infty}$.

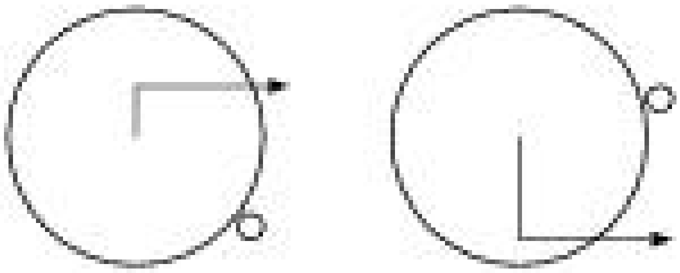

Figure 4: DNA fragments bypassing hurdles using the shorter path. The big ball are DNA fragments, the small balls (which are assumed to be points) are hurdles. The arrows represent the path of the fragment centers.

The Distribution of T_k

To model T_k, we assume that the distribution of hurdles is a spatial Poisson process along the x-axis. In other words, if we follow a DNA fragment along the x-axis, we make the following assumptions (with $P(I)$ being the probability for the fragment bumping into a hurdle in intervall I and $P_{>1}(I)$ being the probability for the fragment bumping more than once into a hurdle in intervall I):

1. We suppose the gel is perfectly homogeneous, so the probability for a certain fragment to bump into a hurdle is everywhere the same: $P([x, x+\Delta x])$ does only depend on Δx but not on x, $\forall x \geq 0$ and $\Delta x \geq 0$.

2. We also suppose that the probability for a certain fragment to bump into a hurdle is independent from where and how many times it bumped into a hurdle earlier: if I and J are disjoint intervals, $P(I)$ and $P(J)$ are independent.

3. Since the gel is very thin, we make the assumption that a fragment can not be in contact with more than one hurdle at the same time. So we have: $P_{>1}([x, x+\Delta x]) = o(\Delta x)$ $\forall x \geq 0$ and $\Delta x \geq 0$.

Under these assumptions, the number of hurdles a fragment meets on its way along the x-axis is a Poisson process with x playing the role of t. So we have:

$$\lim_{\Delta x \to 0} \frac{P([x, x+\Delta x])}{\Delta x} = \lambda,$$

where λ is a real positive parameter. Then the distance between two hurdles is exponentially distributed with parameter λ. We call C the number of hurdles per volume unit. To compute λ, let us imagine a round ball migrating along the x-axis. The cross section of a ball with radius a equals πa^2. Thus, during a short displacement Δx, the swept volume is $\pi a^2 \Delta x$ and the probability that the ball bumps into a hurdle is $\pi a^2 C \Delta x$, hence the simple formula $\lambda = \pi \cdot C a^2$. Since the mass m of the fragment is proportional to its volume, λ is proportional to $Cm^{2/3}$.

As we assume constant and mass-independent speed along the x axis, the T_k are proportional to exponentially distributed random variables with parameter λ. The respective means and variances can now be computed as functions of m: $\forall k > 1$, $\mathbb{E}(T_k) = K_T m^{-2/3}$ and $\mathbb{V}(T_k) = (K_T m^{-2/3})^2$, with K_T being a constant not depending on the fragment i.

The Distribution of S_k

Under quite strong assumption, the modelling of the S_k is easy. Assuming that all fragments bypass the hurdles with the same constant speed in y-direction, we get after a short computation that the S_k are uniformly distributed in the interval $[0, 2K_S m^{1/3}]$, with K_S being a constant that is the same for all fragments. The factor 2 was introduced only for computational reasons.

Definition of τ

We now define for each DNA fragment the integer random variable τ:

$$\tau = \max\{n : \sum_{k=1}^{n-1} (T_k + S_k) < t_0\} \tag{1}$$

Since most of the fragments do not move at all between $t = 0$ and $t = t_0$ (there are much more DNA in the head than in the tail), we assume that the fragments spend much more time bypassing hurdles than migrating. Thus, a given fragment is much more likely to be bypassing an hurdle than to be migrating when the experiment is stopped at time t_0. The sum of the S_k will be much larger than the sum of the T_k and especially for large τ the quantity $\sum_{k=1}^{\tau} T_k$ is a good approximation for time a fragment migrated in the direction of the field.

With this approximation and the constant migration speed v_0 one gets the displacement D of a given fragment as

$$D = v_0 \cdot \sum_{k=1}^{\tau} T_k.$$

The Mean of D

For a given DNA fragment with known mass, the mean of D exists and can be computed similar to Wald's identity, see [10, VII, Theorem 3].

Lemma 3 *Let $(T_k)_{k=1}^{\infty}$ and $(S_k)_{k=1}^{\infty}$ be independent and identically distributed positive random variables with finite mean. Let τ be defined by equation (1). Then*

$$\mathbb{E}(\sum_{k=1}^{\tau} T_k) \ = \ \mathbb{E}(T_1) \cdot \mathbb{E}(\tau)$$

Proof. Since for all $k > 0$, the event $\{\tau \leq k-1\}$ is independent of T_k, and therefore also $\{\tau \geq k\}$, we have:

$$\begin{aligned}
\mathbb{E}(\sum_{k=1}^{\tau} T_k) &= \mathbb{E}(\sum_{k=1}^{\infty} T_k \cdot 1_{[k,+\infty)}(\tau)) \\
&= \sum_{k=1}^{\infty} \mathbb{E}(T_k \cdot 1_{[k,+\infty)}(\tau)) \\
&= \sum_{k=1}^{\infty} \mathbb{E}(T_k) \cdot \mathbb{E}1_{[k,+\infty)}(\tau) \\
&= \mathbb{E}(T_1) \cdot \sum_{k=1}^{\infty} \mathbb{E}(1_{[k,+\infty)}(\tau)) \\
&= \mathbb{E}(T_1) \cdot \mathbb{E}(\tau).
\end{aligned}$$

Note, that the interchange of expectation and infinite sum is justified by the theorem of monotone convergence. □

Corollary 1 *Under the above assumptions the expected displacement is*

$$\mathbb{E}(D) = v_0 \cdot \mathbb{E}(T_1) \cdot \mathbb{E}(\tau) \tag{2}$$

To utilize this result, we need $\mathbb{E}\tau$. Actually, we use an approximation for $\mathbb{E}\tau$ which can easily be computed with good precision.

The definition of τ shows:

$$\begin{aligned}
t_0 &\leq \sum_{k=1}^{\tau} (T_k + S_k) < t_0 + T_\tau + S_\tau \\
t_0 &\leq \mathbb{E}(\sum_{k=1}^{\tau} (T_k + S_k)) < t_0 + \mathbb{E}(T_1 + S_1).
\end{aligned}$$

Applying Wald's identity to the middle term yields lower and upper bounds for $\mathbb{E}\tau$.

$$\frac{t_0}{\mathbb{E}(T_1 + S_1)} \leq \mathbb{E}\tau < \frac{t_0}{\mathbb{E}(T_1 + S_1)} + 1.$$

By introducing this inequality into (2) we get:

$$\frac{v_0 t_0 \mathbb{E}(T_1)}{\mathbb{E}(T_1 + S_1)} \leq \mathbb{E}(D) < \frac{v_0 t_0 \mathbb{E}(T_1)}{\mathbb{E}(T_1 + S_1)} + v_0 \mathbb{E}(T_1).$$

Therefore,

$$\begin{aligned}
\mathbb{E}D &\approx t_0 \cdot v_0 \cdot \frac{\mathbb{E}T_1}{\mathbb{E}T_1 + \mathbb{E}S_1} \\
&= t_0 \cdot v_0 \cdot \frac{K_T m^{-2/3}}{K_T m^{-2/3} + K_S m^{1/3}} \\
&= \frac{1}{K_1 + K_2 m},
\end{aligned} \tag{3}$$

with suitable constants K_1 and K_2. This formula is considered in some references as the best empirical approximation of the (mean) displacement as function of the mass (or the length) [30, 27].

Many other formulae have been proposed in the literature [23, 1, 4]. These formulae are suited to model specific experimental conditions (strength of the electric field, approximate fragment size, gel properties, etc). However, the formula $D = 1/(K_1 + K_2 m)$ and slight modifications seem to prevail.

6 Migration of the Fragment Population

To describe the migration of fragments we use the above point process notation. We assume that the electric field generates a stochastic displacement [11] of the fragments, i.e., each fragment moves independently from the others and from the interaction of the others with the gel. This implies that the random variables D_i are independent and, conditioned on $m_i = m$, identically distributed for every $m > 0$. Further, D_i should be independent from m_j, $j \neq i$. Now we show that these assumptions allow a simple formulation of the migration problem involving a convolution product.

Our basic assumption on the images is that the intensity in one pixel is proportional to the mass of DNA concentrated there. So, we have to consider the mass distribution for the DNA. In point process language, we look for the intensity measures of the process.

The mass intensity measure μ_Ξ [11] is defined as

$$\mu_\Xi(A) \;=\; \mathbb{E}(\sum_{(X,m)\in\Xi} m 1_A(X))$$

for each Borel set A.

In our situation there are two intensity measures: the start intensity μ_X and the end intensity $\mu_{X'}$.

The following assumptions now govern our migration model:

1. The DNA-breaking rate and the DNA-repairing rate are spatially homogeneous. This implies especially that X_i and m_i are independent.

2. The distribution of the displacement D_i of fragment i depends only on its mass m_i and not on X_i. There may be doubts, if this assumption is justified. Indeed, especially when the DNA-concentration is high, fragments may be broken by other

fragments during the migration. However, to get a feasible model, we assume that this effect does not play an important role.

We define f_M as the density of mass, i.e. $\int_{m_1}^{m_2} f_M(m) \cdot \mathrm{d}m$ is the fraction of fragments with mass between m_1 and m_2. Further, f_X denotes the start density, i.e. $\int_{x_1}^{x_2} f_X(x) \cdot \mathrm{d}x$ is the fraction of fragments between $x = x_1$ and $x = x_2$ at the beginning of the electrophoresis, or, to be more precise, the fraction of fragments whose gravity center is between $x = x_1$ and $x = x_2$. Similarly, we define the conditional densities $f_{X'|m}$ and $f_{D|m}$:

$\int_{x_1'}^{x_2'} f_{X'|m}(x) \cdot \mathrm{d}x$ is the fraction of fragments between x_1' and x_2' at the end of the electrophoresis given the mass m, and $\int_{d_1}^{d_2} f_{D|m}(d) \cdot \mathrm{d}d$ is the fraction of fragments with displacement between d_1 and d_2 given the mass m. Further, let f_{μ_X} denote the density of μ_X and $f_{\mu_{X'}}$ the density of $\mu_{X'}$.

Because of the assumption 1, $f_X = f_{\mu_X}$. Finally, let M_c denote the total mass of DNA contained in the considered cell c.

Lemma 4 *With the global density of displacement*

$$f_\sigma(d) \;=\; \frac{1}{M_c} \int_0^\infty m \cdot f_M(m) \cdot (f_D|m)(d) \cdot \mathrm{d}m,$$

we have

$$f_{\mu_{X'}} = f_{\mu_X} * f_\sigma.$$

Proof.

The density $f_{\mu_{X'}}$ of $\mu_{X'}$ can be written as

$$f_{\mu_{X'}}(x') \;=\; \frac{1}{M_c} \int m \cdot f_M(m) \cdot f_{X'|m}(x')\mathrm{d}m.$$

From

$$f_{X'|m}(x') \;=\; \int f_X(x) \cdot f_{D|m}(x' - x)\mathrm{d}x$$

we immediately find our assertion:

$$\begin{aligned} f_{\mu_{X'}}(x') \;&=\; \frac{1}{M_c} \int f_X(x) \cdot \int m \cdot f_M(m) \cdot f_{D|m}(x' - x)\mathrm{d}x\,\mathrm{d}m \\ &=\; \int f_X(x) \cdot f_\sigma(x' - x)\mathrm{d}x. \end{aligned}$$

□

Thus, our model leads to a convolution problem.

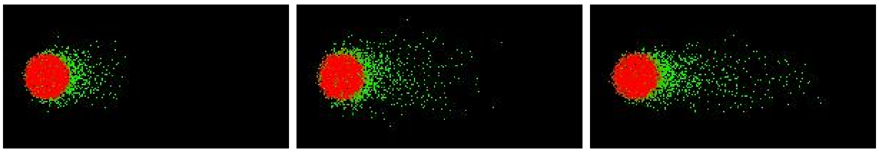

Figure 5: Three simulations from the point process model, parameters top: middle: bottom:

7 Simulation and Comparison to Data

In this section, a simulation is carried out to check qualitatively whether the model for the DNA mass distribution and the model for the DNA migration lead to comet-like shapes. Subsequently, we present a simple method which allows a rough estimate of the model parameters from two specific histograms. These histograms represent, respectively, the horizontal distribution at the beginning and at the end of the electrophoresis. One of these parameters is the required ν_M determining the exponential distribution of fragment masses.

Simulation of the DNA migration

With the software package AntsInFields [15] we implemented the above model, leaving aside the problem of calculating the correct variances. The length of the fragments was sampled from an exponential distribution. The number of fragments was fixed beforehand and assumed to be uniformly distributed over a ball. The distribution of the displacement D was taken as bivariate normal with expectation $(\frac{1}{K_1+K_2 l}, 0)$. The variances σ_x and σ_y were fixed independently from the fragment length m and covariance was assumed to be 0. We stopped the simulations after suitable times to find comet-like shapes.

As Figure 5 indicates, the model is able to capture at least the comet-like shape of the real-world data. The programs written in Oberon are available on request from the last author.

A simple method to estimate the model parameters

The model presented above includes two steps of modelling:

1. the modelling of the distribution of masses as exponential with parameter ν_M:

$$f_M(m) = \nu_M \exp(-\nu_M m)$$

2. the modelling of the dependency between the displacement D and the mass m. The problem of determination of a correct variance formula is ignored: for simplicity the displacement given the mass is assumed to be equal to its mean:

$$D(m) \; = \; \frac{1}{K_1 + K_2 m}.$$

We define a new density f_{μ_M} as follows: $\int_{m_1}^{m_2} f_{\mu_M}(m)\mathrm{d}m$ is the fraction of DNA mass contained in fragments of mass between m_1 and m_2.

As $f_{\mu_M} = m f_M(m)$ one has

$$f_{\mu_M}(m) = \nu_M^2 m \exp(-\nu_M m).$$

Using the convolution lemma 4, it is easy to show that

$$f_\sigma(d) = \frac{\nu_M^2}{d^2 K_2^2}(\frac{1}{d} - K_1)\exp(-\frac{\nu_M}{K_2}(\frac{1}{d} - K_1)).$$

Although this model involves 3 parameters (ν_M, K_1 and K_2), it has only two degrees of freedom, since K_2 and ν_M appear only in the ratio $\frac{K_2}{\nu_M}$. Thus, can only identify the two parameters K_1 and $K = \frac{\nu_M}{K_2}$. Thereto, we need to know the distribution along the x-axis before and after electrophoresis.

Unfortunately, no images of the cells before electrophoresis are available. We only have images of degraded and repaired cells to analyze and images of control cells which have not been grayed. Making the assumption that the DNA distribution in control cells after electrophoresis is similar to the DNA distribution in degraded cells before electrophoresis, we use the images of the control cells to estimate the starting DNA density. This assumption can be justified by the fact that the histograms of control cells are perfectly symmetric, indicating that the DNA fragments in control cells are too big to migrate at all during electrophoresis.

Let us consider two images from the same mouse: an image of a control cell and an image of a degraded cell. Using a JAVA programm, we sum the intensities of all the pixel columns successively, for both images. Thus we obtain discretised estimates of f_{μ_X} and f_{μ_Y}, as depicted in figure 6. Notice that we have aligned the two images arbitrarily. As will become clear later, this causes no problem.

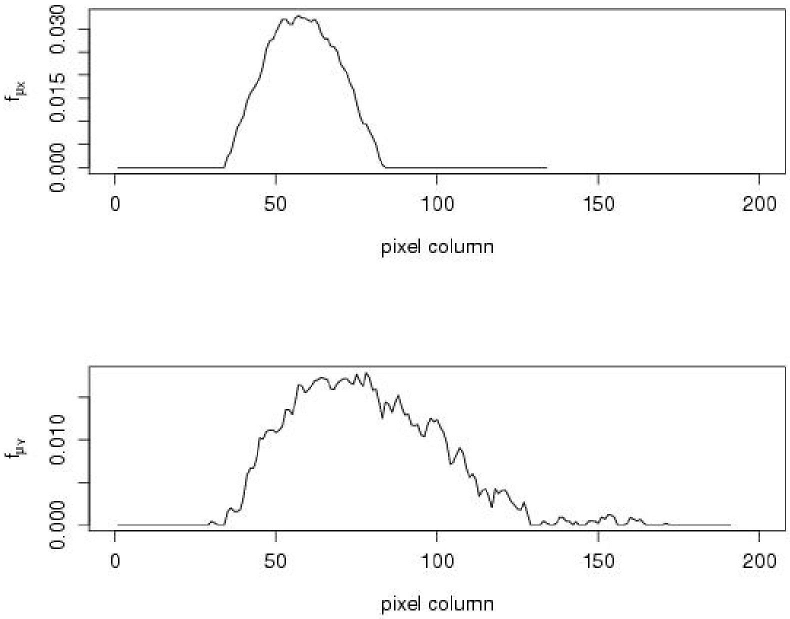

Figure 6: Histograms of a control cell (top) and a damaged cell (bottom) from the same mouse

To estimate the parameters K_1 (whose unit is $pixel^{-1}$) and K, we proceed as follows. For different values of K_1 and K, we perform a discrete convolution of $\hat{f}_{\mu_X}$ and f_σ. Our goal is to find the values for which a certain dissimilarity function between this convolution product and the observed $\hat{f}_{\mu_Y}$ is minimal. Since we do not know the location of the axis origin in the histogram $\hat{f}_{\mu_Y}$, this dissimilarity measure has to be translation invariant. A simple method is to 'subtract' the histogram $\hat{f}_{\mu_X} * f_\sigma$ from the histogram $\hat{f}_{\mu_Y}$ using the criterion of minimal quadratic transportation costs as described in (Boulesteix *et al.*,2003) and to use the variance of the resulting histogram as dissimilarity measure. Clearly, this measure is translation invariant and it is higher for 'very different' histograms than for 'similar' histograms.

To minimize this criterion, we employ the `R` programm `optim` which implements the optimization method of Byrd et al. (1995) and allows to give as inputs lower and upper bounds for each parameter. Here, we set the lower bounds to zero, because the parameters K and K_1 have to be strictly positive. This method yields estimates for K and K_1. A drawback is that our model allows only the estimation of $\frac{\nu_M}{K2}$ and not ν_M, which is actually the parameter we want to estimate. This issue will be addressed in further research.

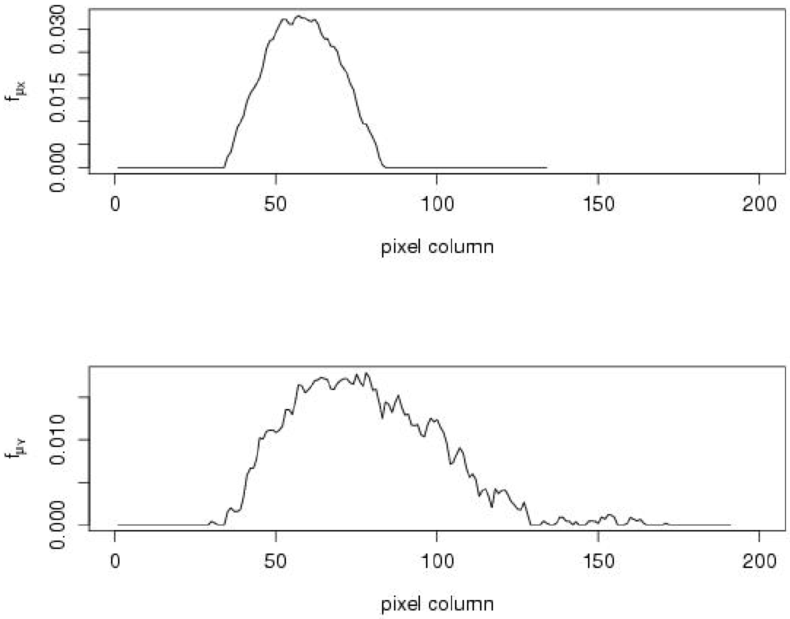

Figure 7: Histograms of a control cell (top) and a damaged cell (bottom) from the same mouse

Results of the parameter estimation

For the two histograms depicted in figure 7, the optimization algorithm yields the following parameter estimates:

$$\hat{K}_1 \approx 0.0074$$

$$\hat{K} \approx 74.$$

For these values, the displacement density f_σ is depicted in figure 8 (left). To evaluate qualitatively the quality of the estimation, we superpose the result of the convolution product of $\hat{f}_{\mu_X} * \hat{f}_\sigma$ obtained with the estimated parameters and the observed $\hat{f}_{\mu_Y}$, as depicted in figure 8 (right). The estimate fits the data well, which indicates that our model is quite realistic.

8 Discussion

In this work, we introduced a stochastic model to describe the comet assay experiment. This model includes two parts. The first part, known in the literature as ’Random Break-

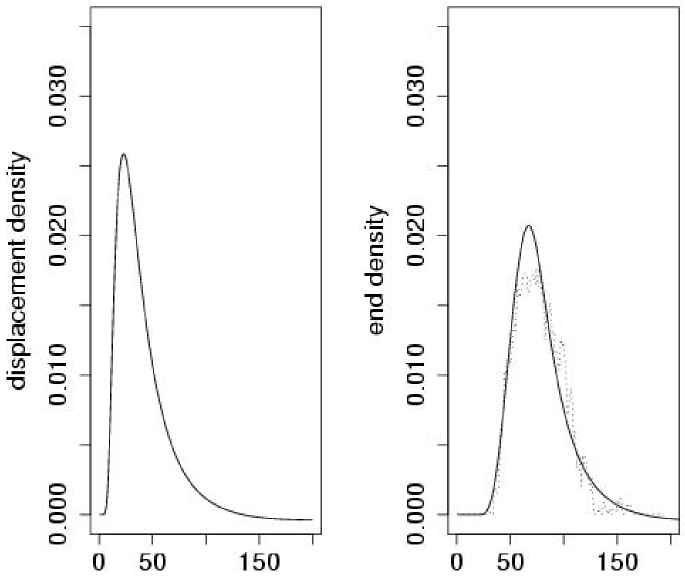

Figure 8: Histogramm of the estimated displacement density with the fitted parameters $K_1 = 0.0074$ and $K = 74$ (left) and of the estimated end density with the fitted parameters (right,solid). On the right panel, the observed end density of the damaged cell is represented as well (dotted), to allow comparison.

age Model' deals with the distribution of length of the DNA fragments. The ultimate goal of this work is to estimate the parameter of this distribution. The second part describes the migration of DNA fragments among gel fibers. This model might be to simple to give a sensible description of the complex mechanisms of DNA damage and electrophoresis. However, it allows mathematical analysis of the obtained cell images, a great advantage compared to more complicated (and unfeasible) theories. Moreover, simulations showed that the model captures phenomenological aspects quite well.

A naive approach to estimate the model parameters is presented in section 7. The major drawback of the present version is, that it allows to estimate the parameter of interest only up to a constant. In future work, this issue should be given much attention. Moreover, the estimation is based on one control cell and one damaged cell, although 30 control cells and 30 damaged cells are available for each mouse. Thus, two major issues should be addressed in future. First, the robustness of the proposed estimation method has to be be studied. Since it is not clear if all cells of the same mouse are equally damaged, the study of robustness might be quite difficult. Second, a criterion is required to address the question of biologists: What is the ability of a given mouse to repair its damaged DNA.

Acknowledgements We thank C. Schindewolf for introduction into the problem, J. Georgii for the development of a Java platform, and F.Friedrich for the simulation via the AntsinFields platform [15].

References

[1] Åkerman B. (1996), Cyclic migration of DNA in gels: DNA stretching and electrophoretic mobility, *Electrophoresis*, **17**, 1027-1036

[2] Alsmeyer G. (1991), *Erneuerungstheorie*, Teubner, Stuttgart

[3] Ashby J., Tinwell H., Lefevre P.A. and Browne M.A (1995), The single cell gel electrophoresis assay for inducet DNA damage (COMET-assay): measurement of tail length and moment, *Mutagenesis*, **10**, 85-90

[4] Bearden J. C. (1979), Electrophoretic mobility of high-molecular-weight double-stranded DNA on agarose gels, *Gene*, **6**, 221–234

[5] Bhattacharya R.N. and Waymire E.C. (1990), *Stochastic processes with applications*, Wiley series in probability and mathematical statistics, Chichester

[6] Böcker W., Bauch T., Müller W.U. and Streffer C. (1997), Image analysis of COMET-assay measurements, *Int.J.Radiat.Biol*, **72**, 449–460

[7] Boulesteix, A.-L, Hösel, V. and Liebscher, V., A comparative study of empirical deconvolution techniques for the COMET-assay. In preparation.

[8] Byrd R.H., Lu P., Nocedal J. and Zhu C. (1995), A limited memory algorithm for bound constrained optimization, *SIAM J.Scientific Computing*, **16**, 1190–1208

[9] Calladine C.R., Collis C.M., Drew H.R. and Mott M.R. (1991), A study of electrophoretic mobility of DNA in agarose and polyacrylamide gels, *J.Mol.Biol.*, **221**, 981–1005

[10] Shiryayev A.N., (1984) Probability, Graduate Texts in Mathematics, Springer-Verlag, New York

[11] Daley D. and Vere-Jones D. (1988), An Introduction to the Theory of Point Processes, Springer Series in Statistics, Springer-Verlag, New York Heidelberg

[12] De Boeck M., Touil N., De Visscher G., Aka Vande P., Kirsch-Volders M. (2000), Validation and implementation of an internal standard in comet assay analysis, *Mutat.Res.*, **469**, 181–197

[13] Fairbairn D.W., Olive P.L. and O'Neill K.L. (1995), The COMET-assay: a comprehensive review, *Mutat.Res.*, **339**, 37–59

[14] Feller (1971), *An introduction to probability theory and its application, Vol.II*, Wiley series in probability and statistics

[15] Friedrich F. (2002), The software package AntsInFields, `http://www.antsinfields.de`

[16] Grimmett G. and Stirzaker D. (1992), *Probability and random processes*, Oxford science publications

[17] Introduction to COMET assay, `http://www.cometassay.com`

[18] Karr A.F. (1986), *Point processes and their statistical inference*, Probability: pure and applied, Manuel Neuts

[19] Kent C.R.H., Eady J.J., Ross G.M. and Steel G.G. (1995), The comet moment as a measure of DNA damage in the COMET-assay, *Int.J.Radiat.Biol.*, **67**, 655–660

[20] Kingman, J.F.C. *Poisson processes*, volume 3 of *Oxford Studies in Probability*. Clarendon Press, Oxford, 1993.

[21] Kraxenberger F., Weber K.J., Friedl A.A., Eckardt-Schupp F., Flentje M., Quicken P. and Kellerer A.M. (1998), DNA double-strand breaks in mammalian cells exposed to γ-rays and very heavy ions, *Radiat.Environ.Biophys.*, **37**, 107–115

[22] Krishnaiah P.R. and Sen P.K. (1984), *Handbook of statistics 4*, Elsevier Science Publishers, Amsterdam.

[23] Lalande M., Noolandi J., Turmel C., Brousseau R., Rousseau J. and Slater G.W. (1988), Bands in gel electrophoresis of DNA, *Nucl.Acids.Res.*, **1988**, 5427–5437

[24] Milligan J.R., Aguillera J.A., Paglinawan R.A., Ward J.F. and Limoli C.L. (2001), DNA strand break yields after post-high LET irradiation incubation incubation with endonuclease-III and evidence for hydroxyl radical clustering, *Int. J. Radiat. Biol.*, **77**,2,155-164

[25] Nelson R. (1995), *Probability, stochastic processes and queuing theory*, Springer-Verlag, Berlin

[26] Ross S.M. (1983), *Stochastic processes*, Wiley series in probability and mathematical statistics, Chichester.

[27] Serwer P. (1989), Sieving of double-stranded DNA during agarose gel electrophoresis, *Electrophoresis*, **10**, 327–331

[28] Slater G. W., Mayer P. and Drouin G. (1996), Migration of DNA through gels, *Methods Enzymol.*, **270**, 272–295

[29] Software for Comet Analysis: VISCOMET, TILL Photonics

[30] Southern E.M. (1979), Measurement of DNA Length by Gel Electrophoresis, *Anal.Biochem.* **100**, 319–323