Appl. Math. and Comp. Sci., 1994, vol.4, No.2, 163–177

RAPID GENOME EVOLUTION, CANCER AND HERITABLE DISEASES: A MODELLING PERSPECTIVE

MAREK KIMMEL*

Rapid changes in composition and amount of human DNA are accompanying diseases like cancer or certain heritable disorders. Dynamics of these changes can be better understood if viewed using probabilistic models. The agreement between the mathematical model and the experimental data lends credence to the biological theories devised to explain experimental observations.

1. Introduction

The amount of DNA per cell remains constant from one generation to another because during each cell cycle the entire content of DNA is duplicated and then at each mitotic cell division the DNA is evenly apportioned to two daughter cells. However, recent experimental evidence shows that for a fraction of DNA, its amount per cell and its structure undergo continuous change. In this paper we attempt a brief review of circumstances in which this occurs, focused on the connections between genomic changes, cancer and hereditary diseases. Our main purpose is to demonstrate how stochastic modelling may illuminate the mechanisms involved.

One way the genome of cancer cells may rapidly evolve is by an increase in gene copy number, referred to as *gene amplification*. Gene amplification can be enhanced by conditions that interfere with DNA synthesis and is increased in some mutant and tumour cells. Increased number of genes may produce an increased amount of gene products and, in tumour cells, confer resistance to chemotherapeutic drugs. Amplification of oncogenes has been observed in many human tumour cells and also may confer a growth advantage on cells which overproduce the oncogene products (for an overview see survey by Stark (1993) and Windle and Wahl (1992)).

In the classical experiments of Schimke and his coworkers (Brown *et al.*, 1981; Kaufman *et al.*, 1981), the anticancer drugs served to select cells with amplified genes. In part of cell lines, when the selective agent was removed, the cells with amplified genes were gradually disappearing from the population. The stochastic mechanism leading to this reversal is discussed in Section 2. It was observed that in such cases the amplified genes were located on extrachromosomal fragments of DNA called *Double Minute Chromosomes (DM's)*. In other cases, the amplification was stable, i.e. persisted after the selective agent had been removed. In such cases, the amplified genes are usually located on elongated chromosome arms. The most regular of these structures exhibit a regular band structure (the so-called *Homogeneously Staining Regions* or *HSR's*), but other less regular structures are also observed. They are either

* Department of Statistics, Rice University, Houston, TX 77251, USA

caused by reintegration of extrachromosomal genes as proposed by Wahl (Windle *et al.*, 1991), or they arise by a separate mechanism as proposed by Stark (Smith *et al.*, 1992). Mathematical models show that depending on circumstances each of the two variants of stable amplification is plausible (Sections 3 and 4).

One of the questions related to gene amplification concerns the so-called *primary event*, during which the first additional copy of the gene appears in a single cell, which then gives origin to the "amplified" clone. Is the primary event spontaneous or is it induced by the same agent which selects the amplified phenotype? Attempts at answering this question were made (Tlsty *et al.*, 1989) using the classical *fluctuation analysis* method of Luria and Delbrück (1943). Again, stochastic modelling allows us to improve upon these considerations (Section 5).

The impact of drug resistance on cancer treatment is of considerable practical importance. Within the framework of gene amplification, this question was examined by Harnevo and Agur (1991, 1992, 1993) (Section 6).

Recently, it has been discovered that a number of hereditary diseases are caused by dynamic expansions of chromosome regions including short multiple *tandem repeats* of DNA (Richards and Sutherland, 1992). Similar hypothesis was advanced for the most common form of colon cancer (Marx, 1993). Evolution of repetitive DNA sequences can be viewed as the action of a discrete stochastic dynamical system and it can be modelled in this way (Section 7). In certain hereditary diseases, initial modest expansion is followed by explosive "proliferation" of DNA repeats (Section 8). These problems are a subject of active research.

2. Unstable Gene Amplification

In some populations of cells with double minute chromosomes, both the increased drug resistance and the increase in number of gene copies are *reversible*. The classical experiment confirming this includes transfering the resistant cell line into drug-free medium (Brown *et al.*, 1981; Kaufman *et al.*, 1981), where the cells gradually lose resistance to the drug by losing extra gene copies.

The population distribution of numbers of gene copies per cell can be estimated by flow cytometry after staining gene products. In the experiments mentioned (Brown *et al.*, 1981; Kaufman *et al.*, 1981), two features of these distributions are notable:

1) As expected, the proportions of resistant cells (with amplified genes) decrease with time.
2) Less obvious, the shape of the distribution of the number of gene copies limited to the resistant cell subpopulation seems to remain stable during the loss of resistance.

A mathematical model of the unstable drug resistance should take into account i) stochastic changes in number of gene copies from one generation to another and ii) the stochastic variability in cell lifetimes. One stochastic process which accomodates both i) and ii) is the random walk superimposed on the time-continuous branching process of cell proliferation, i.e. a *branching random walk* (Kimmel and Stivers, 1994). We consider a population of abstract particles of types $j = 0, 1, 2, \ldots$:

1. The lifespans of all particles are independent identically distributed exponential random variables with mean $1/\lambda$.

2. At the moment of death, a particle of type $j \geq 1$ produces two progeny particles each belonging to type $j+1$ with probability b, to type $j-1$ with probability d, and to type j with probability $1-b-d$. A particle of type $j=0$ produces two progeny of type 0.
3. The process is initiated at time $t=0$ by a single particle of given type i.

The simplest models of gene amplification in (Kimmel and Axelrod, 1990) and (Kimmel and Stivers, 1994) assume the above process (Fig. 1). Cells with 2^{j-1} gene copies are said to belong to type j (with 0 gene copies, to type 0). The parameters b and d are the probabilities of gene *amplification* and *deamplification*, respectively.

One of the properties of Markov processes with absorbing states is the possibility of existence of the quasi-stationary distributions. In intuitive terms, the unabsorbed part of the probability mass of the process, while constantly shrinking, approaches a limit if it is properly normed. The Yaglom theorem for subcritical branching processes (Athreya and Ney, 1972) can be quoted as an example. It is this property that explains the apparent stability of distributions of gene copy number per cell in the resistant subpopulation, placed in the non-selective medium.

The numerical values of the probabilities of gene amplification and deamplification can be estimated based on data in (Brown *et al.*, 1981; Kaufman *et al.*, 1981). The probabilities of deamplification (d_i) are of the order of 0.10 in both cases, while the probabilities of amplification (b_i) are about 5 times lower. The process is strongly subcritical.

More realistic models of unstable amplification are discussed in (Kimmel and Axelrod, 1990) and (Kimmel and Stivers, 1994). They yield qualitatively similar results.

The classical explanation for the loss of resistance in cells with amplified DNA in extrachromosomal elements is that in the absence of selective pressure cells with extra gene copies grow slower and are outgrown by the sensitive cells (Kaufman *et al.*, 1981). Our model assumes a purely stochastic mechanism. A combination of two mechanisms is likely. For further comments, see (Kimmel and Axelrod, 1990).

3. Reintegration

In the experimental system of Windle, Wahl and co-workers (Windle *et al.*, 1991) amplified genes residing on extrachromosomal elements were observed in cell cultures 8–9 generations old, while predominantly chromosomally amplified genes were seen after about 30 generations (only these two time points were investigated). This can be interpreted as an indication that extrachromosomal genes are reintegrated into chromosomes. Fitting these latter data requires a different mathematical model (Kimmel *et al.*, 1992).

In this model, the basic indivisible unit which serves as the template for the production of additional gene copies is the *amplicon*, which contains at least one copy of the target gene. The size of such structures could range from submicroscopic to an entire arm of a chromosome and they may be circular or linear. The *acentric (replicating) element (ARE)* is understood to be an extrachromosomal molecular structure containing one or more amplicons but no centromere. The *reintegrated element (RE)* is the ARE after it has reintegrated into a chromosome.

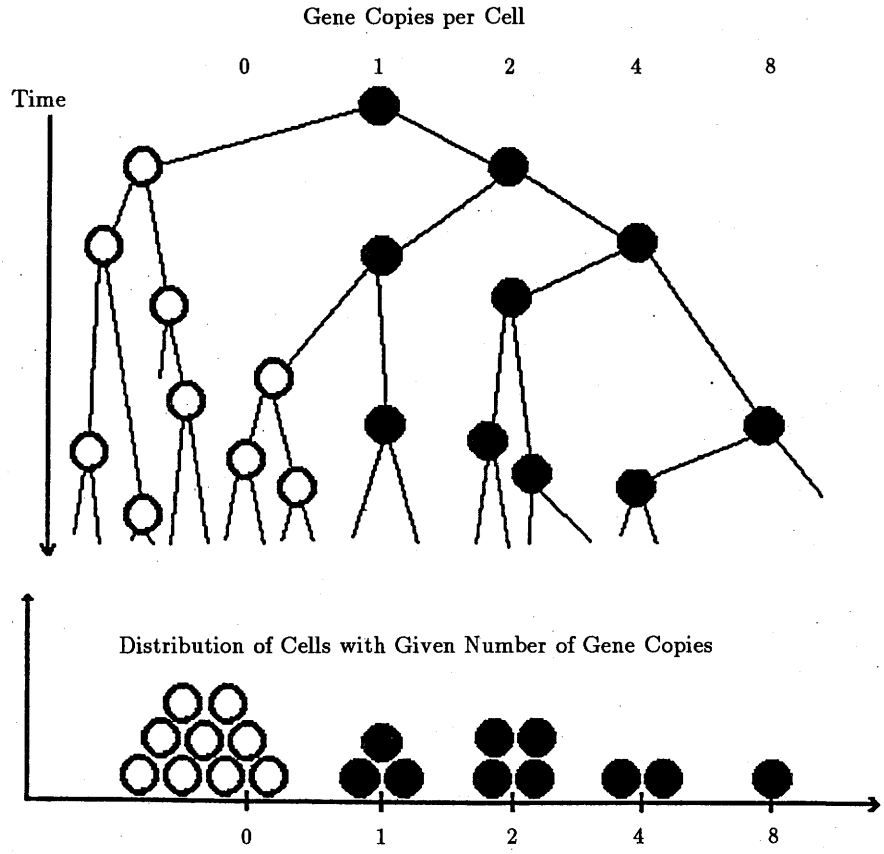

Fig. 1. Gene amplification and deamplification in a representative small cell pedigree of cells grown under nonselective conditions, according to the branching random walk model. Circles represent cells with the number of gene copies per cell indicated across the top row. *Open circles*, cells with no gene copies; *solid circles*, cells with one or more gene copies. Each cell with at least one gene copy can give rise to progeny cells that have double the number of gene copies, with probability b, half that number, with probability d, or the same number, with probability $1-b-d$. The histogram at the bottom shows the resulting distribution of gene copies per cell after time t.

The following processes are considered in the model: i) change in the number of ARE's per cell, ii) change in the number of amplicons per ARE, and iii) reintegration of acentric elements into chromosomes.

The dynamics of these processes are based on the following rules (Fig. 2):

1. All acentric elements evolve independently of each other.
2. Types of elements: i) acentric elements containing $i = 1, 2, \ldots$ amplicons, and ii) chromosomes with one or more sites containing reintegrated elements each containing $i = 1, 2, \ldots$ amplicons.

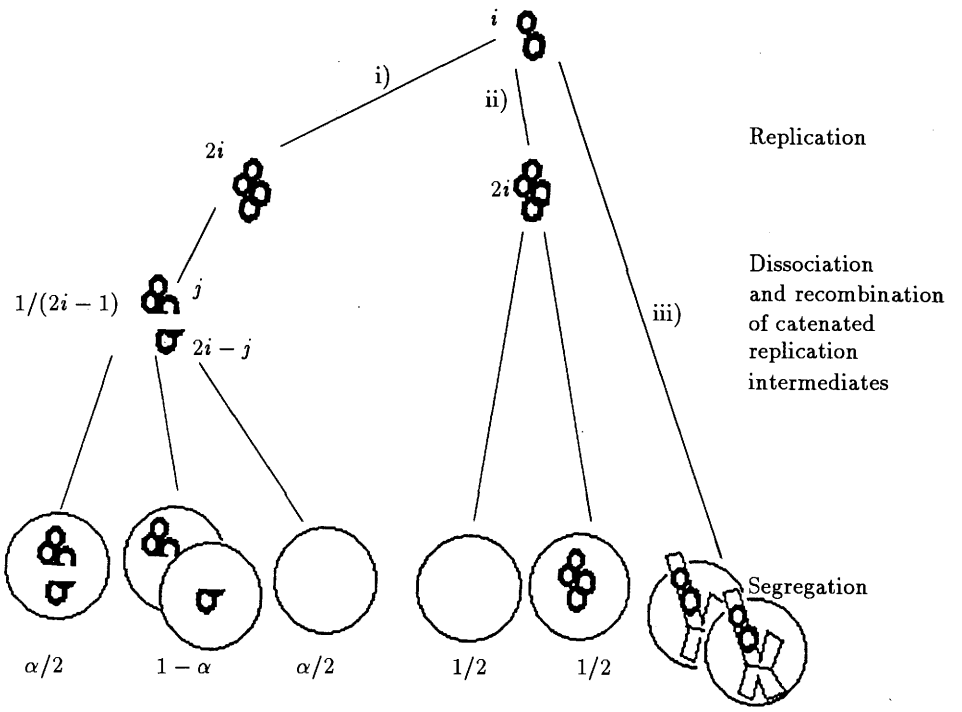

Fig. 2. Schematic representation of the events in the gene amplification model. The process is initiated when a DNA fragment containing a DHFR gene is deleted from its site on a chromosome. This extrachromosomal acentric DNA fragment shown on the top, and each of its descendants, can follow one of three pathways in each cell generation. i) It may replicate and break unevenly into two pieces which randomly segregate into daughter cells. ii) It may replicate and not break before random segregation. These steps are reiterated producing further extrachromosomal elements with different numbers of gene copies per fragment. iii) Each of these elements may reintegrate into a chromosome were they are stably replicated and segregated. The illustration shows one extrachromosomal element containing i gene copies, where i may be $1, 2, \ldots$, etc. Each extrachromosomal element, and each of its descendants can independently follow these pathways. Further explanation of the symbols is given in the text.

3. In each cell generation, three types of processes can occur for each acentric extrachromosomal element: i) ARE replicates, replication products dissociate (this process involves topological resolution of the intertwined products of DNA replication), and the two replication products segregate independently, ii) ARE replicates and replication products do not dissociate, but coexist as a dimer or higher multimer, iii) one or more ARE(s) reintegrate into a centric chromosome (or fragment).

4. With probability a, the ARE containing i amplicons replicates to yield a product with $2i$ amplicon copies. The catenated replication product then dissociates producing two acentric molecules. This process, possibly followed by recombination between the acentric molecules, results in a pair of molecules containing, respectively, j and $2i-j$ amplicons, where $j = 1, \ldots, 2i-1$. It is assumed that the probability of each pair $(j, 2i-j)$ is the same, equal to $1/(2i-1)$. The molecules segregate so that they both go to the same daughter cell with probability α, and go to different daughter cells with probability $1-\alpha$.
5. With probability b, the ARE with i amplicon copies replicates to yield a product with $2i$ amplicon copies, but this replication product does not dissociate. It then goes with equal probability to one of the two daughters.
6. With probability c, the ARE containing i copies of the amplicon, integrates into a chromosome with a centromere and then replicates and segregates with the chromosome. This results in each daughter cell containing an equal number of RE copies. Further increases or decreases in gene copy number are envisioned to occur at extremely low rate at this site at subsequent cell divisions. The probability of reintegration is $c = 1-(a+b)$.

If we consider a randomly selected cell lineage, we may formally define the following random variables:

- $X_n^i(\omega)$, the number of acentric elements with i copies of the amplicon, in the n–th cell generation,
- $Y_n^i(\omega)$, the number of elements reintegrated into chromosomes, with i copies of the amplicon, in the n–th cell generation.

The sequence $\{\{(X_n^1, Y_n^1), (X_n^2, Y_n^2), \ldots\}, n = 0, 1, 2, \ldots\}$, is a *multitype Galton-Watson process with a denumerable infinite number of particle types.*

The following are the general consequences of the model assumptions. Mathematical proofs and derivations are contained in the Appendix to (Kimmel *et al.*, 1992).

1. Among cells with at least one acentric element copy, there will be initial increase in number of acentric elements per cell, and the number of amplicon copies per acentric element. Subsequently, as the acentric elements become reintegrated, their number per cell will decrease and the proportion of cells with stably integrated copies will increase.
2. An eventual consequence will be a population of cells containing only integrated elements with a spectrum of amplicon copy numbers at one or more chromosomal locations. The model enables computation of this distribution, at different values of a, b, c, and α.

The model described here fits the experimental data in (Windle *et al.*, 1991). It may find broad applicability and may help to understand the complicated kinetics and the multiple structures observed in many situations in which genes are amplified only transiently in an extrachromosomal state.

4. Unequal Chromatid Exchange

Amplification of the CAD gene in Syrian hamster BHK cells growing in the presence of the inhibitor PALA has been described in (Smith *et al.*, 1992). The authors of that work have observed amplified genes associated with very regularly repeated structures on chromosomes without observing extrachromosomal elements containing amplified genes. They reported a broad distribution of the number of repeats per chromosome. These authors (Smith *et al.*, 1992) propose a biological model to explain gene amplification in this system which is different from the biological model proposed by Windle and co-workers (Windle *et al.*, 1991); specifically they propose the generation of chromosomally located tandem arrays without intervention of any extrachromosomal intermediates. The biological model in (Smith *et al.*, 1992) has three major phases:

- First, an initial gene duplication is generated by recombination between telomeres and centromeres of sister chromatids as suggested in (Smith *et al.*, 1992). Telomeres and centromeres are noncoding repeated sequences at the ends and in the centers of chromosomes, respectively.
- Second, long tandem arrays of genes are produced by repeated unequal sister chromatid exchanges (misalignment and recombination).
- Third, condensed structures containing amplified genes are generated by an undetermined mechanism.

We modelled the second stage of gene amplification, resulting from repeated unequal chromatid exchange (i.e. misalignment followed by recombination). The model is based on the following principles (Fig. 3.):

1. A cell of generation m contains X_m repeats on each sister chromatid. It is assumed that $X_0 = 2$.
2. If the chromatid contains $X_m = k$ repeats, then the misalignment by S_m repeat units has a symmetrized geometric distribution with parameter p. That is,
$$P(S_m = s|X_m = k) = p^{|s|}/2(1 + p + p^2 + \ldots + p^{k-1})$$
if $s = -(k-1), \ldots, -1, 1, \ldots, k-1$, and
$$P(S_m = 0|X_m = k) = 1 - \sum_{s \neq 0} P(S_m = s|X_m = k)$$
3. If the chromatid contains $X_m = k$ repeats and the misalignment is $S_m = s$ then the number of crossover sites, N_m, is a Poisson random variable (Haldane's model) with parameter $(k-s)\mu$ proportional to the length of the paired region. The chromatid exchange is effective only if the number of crossovers is odd.
4. As a consequence of the s–unit misalignment and recombination, the number of repeats located on the first sister chromatid increases to $k+s$ and the number of repeats on the second sister chromatid decreases to $k-s$.
5. After cell division, sister chromatids segregate and one daughter cell receives a chromatid with $k+s$ repeats, and the other with $k-s$ repeats.

The sequence $\{X_m, m = 0, 1, \ldots\}$ is a Markov process describing the evolution of the number of repeats in successive cell generations.

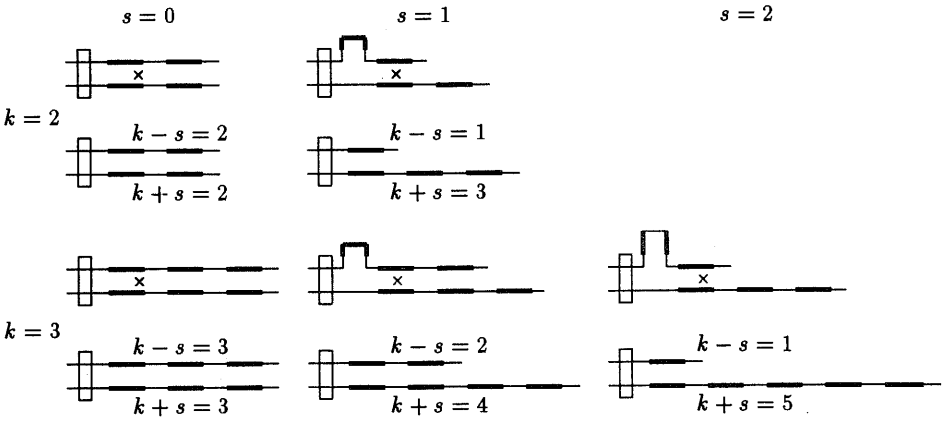

Fig. 3. Mathematical model of gene amplification by unequal sister chromatid exchange. The diagrams depict repeat units of regions containing the CAD gene (thick lines) on chromatids. A pair of sister chromatids is attached to a common centromere (rectangle). Two situations are illustrated. The first, for a cell with two gene copies on each chromatid ($k = 2$), and the other, for a cell with three gene copies ($k = 3$). Analogous diagrams for $k \geq 4$ are not shown. In the case of $k = 2$, there are two possible pairings of chromatids. If there is no misalignment ($s = 0$), then recombination yields two chromatids with two gene copies each ($k - s = k + s = 2$). If there is misalignment by one repeat unit ($s = 1$), then recombination may yield one chromatid with one gene copy ($k - s = 1$) and a sister chromatid with three gene copies ($k + s = 3$). In the case of $k = 3$, there are three possible pairings of chromatids. If there is no misalignment ($s = 0$), then recombination yields two chromatids with three gene copies each ($k - s = k + s = 3$). If there is misalignment by one repeat unit ($s = 1$), then recombination may yield one chromatid with two gene copies ($k - s = 2$) and a sister chromatid with four gene copies ($k + s = 4$). If there is misalignment by two repeat units ($s = 2$), then recombination may yield one chromatid with one gene copy ($k - s = 1$) and a sister chromatid with five gene copies ($k + s = 5$). Crossover is denoted by an X. At mitosis, sister chromatids segregate into sister cells; one cell receives a chromatid with $k + s$ repeats and the other with $k - s$ repeats. In PALA containing medium, cells with gene copy number below a threshold die, while remaining cells proliferate and may reiterate misalignment and recombination.

The parameter p is related to the extent of misalignment. It is defined as the factor by which the probability of misalignment by $s + 1$ units decreases compared to the probability of misalignment by s units. For example if p were equal to 1 and there were k repeat units on each sister chromatid, then misalignment by $s = 0, 1, 2, \ldots, k - 1$ would be equally probable. On the other hand, if p were equal to 0.5 and there were k repeat units on each sister chromatid, then misalignment by $s = 0, 1, 2, \ldots, k - 1$ would have probabilities $a, a/2, a/4, \ldots, a/2^{k-1}$, where $a = 1/2(1 - 2^{-k})$ is a norming factor.

Normal cells which have not amplified their CAD gene do not produce progeny when grown in selective conditions, namely PALA-containing medium. We explored two variants, i) that only cells with a single copy of the gene are eliminated in the selective conditions, and ii) that all cells with a single copy of the gene, and a fraction

d of cells with two gene copies, are eliminated. Cells with more than two gene copies are not eliminated. Variant ii) provided an acceptable fit to experimental data.

An interesting feature of this model is the behaviour of the number of repeat units under non-selective conditions. Despite the fact that the mean number of repeat units is conserved from one generation to another, the absorbing state corresponding to a single repeat unit is reached with probability one. This behaviour is analogous to that displayed by critical branching processes. (see Athreya and Ney, 1972). Biologically this means that in non-selective conditions the cells with amplified sequences gradually disappear from the population even if they are not at a growth disadvantage; although rare cells with a very large number of amplified sequences might exist.

5. In Search of the Primary Event

The determination of mutation rates is an important experimental procedure for characterizing mutation processes. The accepted method of determining mutation rates, the fluctuation test, was introduced by Luria and Delbrück (1943) and used for nearly 50 years without major modification. The protocol and analysis are based on a view of sudden and inherited changes in phenotype being due to single events which are effectively irreversible.

It has become appearent that some inherited changes in phenotype are due to more than one genetic change and that, in some cases, the changes are reversible at non-negligible rates. Examples include gene amplification, multistage carcinogenesis, radiation and chemical DNA damage and repair. Molecular analysis has revealed several stages in the process of gene amplification, some of which are reversible (Windle *et al.*, 1991).

A required refinement of the Luria-Delbrück model includes at least two-stage mutation with the first stage reversible (Kimmel and Axelrod, 1994). The hypotheses are as follows (Fig. 4):

1. Two types of cells exist in the population: *type 0* nonmutant cells and *type 1* mutant cells.
2. All cells in the population have interdivision times equal to $\ln 2$.
3. Each cell, at the moment of division, gives birth to two daughter cells. The type of each of these daughters is the same as that of the mother cell.
4. Following division, a *type 0* daughter cell undergoes transformation into a *type 1* cell, with probability α_{01}; a *type 1* daughter cell undergoes a reverse transformation into a *type 0* cell, with probability α_{10}; a *type 1* daughter cell undergoes irreversible transformation into a *type 2* cell, with probability α_{12}.

Given parameter values, the model predicts the distribution of the number of non-mutant and mutant cells at time t in a population started at time 0 by a single non-mutant cell. In particular, the following *observable* variables are of interest:

- $N(t)$, the expected total number of non-mutant and mutant cells at time t,
- $r(t)$, the expected number of mutant cells at time t,
- $P_0(t)$, the probability of mutant cells being absent from the population at time t.

Given experimental values of $N(t)$, $r(t)$, and $P_0(t)$, it is possible to estimate the mutation rates and probabilities.

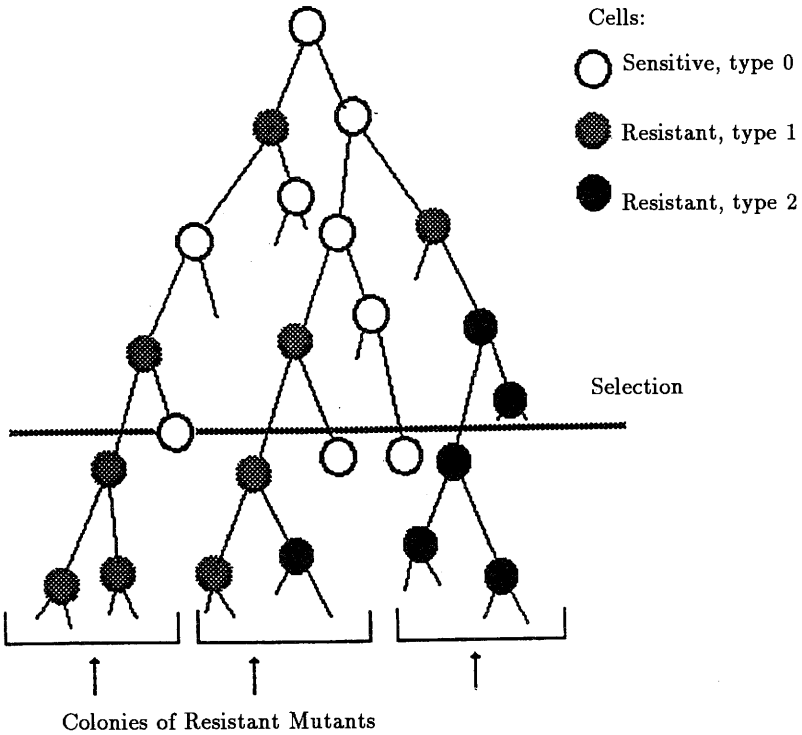

Fig. 4. Schematics of transitions assumed in the two-stage model. Specific hypotheses spelled out in the text.

Comparison of the estimates of mutation rates by the single and two stage models is interesting. The estimates of the single mutation rate and the first forward mutation rate are both low (10^{-6} to 10^{-7}). However, the two stage model indicates high rates of the reverse mutation and of the second forward mutation (0.05 to 0.85, correspondingly). These high probabilities are similar to estimates of probabilities of gene amplification and deamplification that we previously estimated by other methods (cf. Section 2).

Estimates of mutation rates have been used to compare different cell lines. For instance, study (Tlsty *et al.*, 1989) concluded that tumourigenic cell line WB_{20} had a rate of gene amplification that was tenfold greater than that of the non-tumourigenic cell line GN_5. In the two stage model the probability of the first forward step, α_{01}, is tenfold greater for the tumourigenic cell line than the non-tumourigenic cell line.

6. Prospects for Chemotherapy of Cancer

Resistance to antineoplastic drugs has been a major impediment to the successful treatment of cancer. Recent studies suggest that several mechanisms are responsible for the emergence of drug resistance but that high levels of resistance and poor prognosis are strongly associated with gene or oncogene amplification.

In recent years the problem of drug resistance in cancer has been mathematically attacked by many authors, including Coldman and Goldie (1983) (for an overview see the book by Wheldon (1988)). Underlying these models was the assumption that drug resistance in cancer results from a single mutational event whose probability is constant and independent of external constraints.

Harnevo and Agur (1992) introduce a model which treats the emergence of drug resistance as a dynamic process rather than a single event. Using this model, based on their previous works (Harnevo and Agur, 1991), they focus on gene amplification as one of the mechanisms that may lead to drug resistance, and show how changes in the underlying assumptions affect the predictions about treatment efficacy. The mathematical modelling results suggest that under gene amplification dynamics with high amplification probability, protocols involving frequent low-concentration dosing may result in the rapid evolution of large fully resistant residual tumours; the same total doses divided into high-concentration doses applied at larger intervals may result in partial or complete remission.

Another suggestion is that treatment prognosis may be largely improved if cells bearing a large gene copy number suffer high mortality. Therefore, it may be interesting to examine the possibility of incorporating in the treatment an agent (hypothetical, at present) that increases the mortality of cells carrying highly amplified genomes.

7. Evolution of Tandem Repeats

The origin and evolutionary function of repetitive DNA in eukaryotic cells is still not fully understood. This problem is a part of the more general *DNA C-value paradox*, i.e. the observation that most eukaryotes, unlike bacteria, contain large amounts of non-coding DNA (Cavalier and Smith, 1985).

The interest in repetitive DNA recently increased for two reasons. One reason is that because of its variable nature it provides convenient markers for the analysis of human genetic linkage. The other is that regions of variable repeats have been recently implicated in the origin of several heritable diseases and cancer.

One of the primary candidates for the mechanism of creation and propagation of repeated DNA is the homologous or nonhomologous recombination between sister or nonsister chromatids during mitosis or meiosis. It seems quite obvious that if repeats or quasi-repeats already exist, unequal chromatid exchange propagates them. This mechanism was exploited in simulations and mathematical models, (cf. Axelrod *et al.*, 1994). What seems less obvious is that unequal chromatid exchange, even in the presence of mutations, can *generate* repeats and quasi-repeats. This has been demonstrated by Smith (1976) who showed that "DNA whose sequence is not maintained by selection will develop periodicities as a result of random crossover" and then confirmed and deepened by Stephan (1989). One of the more recent studies is (Harding *et al.*, 1992).

The following is based on (Baggerly and Kimmel, 1994). Initially, a sequence A of base pairs A, C, G and T represented by numbers between 0 and 3 is randomly generated. An exact copy of this sequence, A', is then constructed. These two

sequences are treated as sister chromatids. The sequences are subjected to a series of mutations, and misalignments followed by attempted crossovers, in a series of cycles. The attempted crossover succeeds only if there exists a region of homology of a given length m between the two misaligned sequences.

As an example with parameter $m = 4$, let us consider:

A ....012322310123....

A' ....012322310123....

A random misalignment amount is determined (say -4), so A now misaligns four elements to the left, and a crossover area (designated with x's) is chosen

A ....012322310123....
××××
A' ....012322310123....

The sequence 2310 is not the same as 1232, so there is not enough homology for the crossover to occur. Subsequently, A and A' are realigned, and then displaced by another random amount (say 8) so A now misaligns by eight elements to the right, and a new crossover region is determined.

A .....012322310123....
××××
A' .....012322310123....

Here crossover does occur, and the resultant sequences are

$A(new)$ ....0123....

$A'(new)$ ....01232231012322310123....

so that $A(new)$ is eight elements shorter than the previous A.

We then replicate $A(new)$ and repeat the process of misalignment and attempted crossover. After some specified number of crossover attempts (whether or not they result in success), a mutation occurs – one element in the sequence A is randomly chosen and replaced with a random element (0, 1, 2 or 3).

A simulation study based on these principles indicates that the average number of crossovers per cycle levels off. What is happening is that the sequences are each becoming dominated by a repeated subsequence as is the case in Smith's model, but in addition it is apparent that the lengths and patterns of these selfsame subsequences follow well-defined distribution. The stability of the distributions has a stochastic character, since mutations are still present. One manifestation of this is the occasional "quantum shift" from one repeating subunit of a given length to a similar one of the same length.

Determination of the composition of these distributions is a nontrivial problem because of the great number of possible repeat patterns.

Another related question is the evolution of DNA structures more complicated and meaningful than simple repeats. An attempt at analysis based on mathematical linguistics has been published by Searls (1992).

8. Tandem Repeats in Heritable Diseases and Cancer

Interspersed repeated DNA sequences located close to or within genes have brought the mechanism of their evolution into the area of human molecular genetics. These sequences have a unique form of mutation: variation in copy number. The rate of mutation is related to the copy number and therefore the mutability of the product of a change in copy number is different from that of its ancestor. These *dynamic mutations* are responsible for at least five human genetic diseases, the first two discovered being the fragile X syndrome and myotonic dystrophy (see the survey by Richards and Sutherland (1992)):

- The fragile X syndrome, caused by a mutation of the FMR–1 gene characterized by expansion of the $(CCG)_n$ repeats (normal 6–60, carrier, 60–200, affected >200 repeats).
- Myotonic dystrophy, caused by a mutation of the DM–1 autosomal gene characterized by expansion of the $(AGC)_n$ repeats (normal 5–27, affected >50 repeats).

These two inherited human diseases had previously been distingushed by two notable features, anticipation (DM) and the Sherman paradox (FMR). The Sherman paradox is that symptoms become more extreme in subsequent generations (Fu *et al.*, 1992). Anticipation is the earlier onset of symptoms and increased severity in subsequent generations (Redman *et al.*, 1993). These features have recently been correlated with changes in DNA. In each case a a trinucleotide represented a few times in unaffected parents is found in multiple tandem copies in progeny. The number of tandem copies is increased dramatically (× 10–100) in affected individuals. The number of repeat sequences has been correlated with the time of onset and the severity of symptoms.

Several important questions that have not been answered are:

1. What is the mechanism of relative stability of the number of repeat sequences in normal people (not in affected families)?
2. What is the mechanism of the modest increase in repeat sequences in unaffected carriers?
3. What is the mechanism of the rapid expansion of the number of repeat sequences in affected progeny within one or two generations.

There are two classes of mechanisms which could account for the observed dynamics of increase in sequence number. One is a threshold mechanism which requires an initial modest expansion and then, when the threshold is exceeded, causes a subsequent rapid expansion. The second mechanism is a uniform non-linear process which stays relatively stable when the repeat count is low but accelerates after the initial modest increase. Mathematical models may be based on the theory of branching processes.

Recently, a seemingly similar phenomenon of "proliferation of repeats" was discovered in one of the forms of human colon cancer (Marx, 1993).

9. Concluding Remarks

Probabilistic models reproduce the dynamics of rapid changes in the DNA. By comparison with experimental data, they allow us to estimate the values of otherwise non-observable parameters. Even more important, they help to verify consistency and feasibility of biological theories put forward as explanations of experimental observations.

Acknowledgement.

The author was supported by the NSF Grant DMS 9203436. Input from David Axelrod is gratefully acknowledged.

References

Athreya K.B. and Ney P.E. (1972): *Branching Processes.* — New York: Springer.

Axelrod D.E., Baggerly K.A. and Kimmel M. (1994): *Gene amplification by unequal chromatid exchange: Probabilistic modelling and analysis of drug resistance data.* — J. Theor. Biol., (to appear).

Baggerly K.A. and Kimmel M. (1994): *Emergence of stable DNA repeats from random sequences under unequal sister chromatid exchange.* — Proc. World Congress *Nonlin. Analysts*, Tampa, Florida, August 1992, (to appear).

Brown P.C., Beverly S.M. and Schimke R.T. (1981): *Relationship of amplified Dihydrofolate Reductase genes to double minute chromosomes in unstably resistant mouse fibroblasts cell lines.* — Mol. Cell. Biol. v.1, pp. 1077–1083.

Cavalier-Smith T. (1985): *Introduction: the evolutionary significance of genome size.* — In: T. Cavalier-Smith (Ed.), The Evolution of Genome Size, New York: Wiley, pp.1–36.

Coldman A.G and Goldie J.H. (1983): *A model for the resistance of tumour cells to cancer chemotherapeutic agents.* — Math. Biosci., v.65, pp.291.

Harding R.M., Boyce A.J. and Clegg J.B. (1992): *The evolution of tandemly repetitive DNA: recombination rules.* — Genetics, v.132, pp.847–859.

Harnevo L.E. and Agur Z. (1991): *The dynamics of gene amplification described as a multitype compartmental model and as a branching process.* — Math. Biosci., v.103, pp.115–138.

Harnevo L.E. and Agur Z. (1992): *Drug resistance as a dynamic process in a model for multistep gene amplification under various levels of selection stringency.* — Cancer Chemother. Pharmacol., v.30, pp.469–476.

Harnevo L.E. and Agur Z. (1993): *Use of mathematical models for understanding the dynamics of gene amplification.* — Mutat. Res., v.292, pp.17–24.

KaufmanR.J., Brown P.C and Schimke R.T. (1981): *Loss and stabilization of amplified dihydrofolate reductase genes in mouse sarcoma S–180 cell lines*, Mol. Cell. Biol., v.1, pp.1084–1093.

Kimmel M. and Axelrod D.E. (1990): *Mathematical models of gene amplification with applications to cellular drug resistance and tumourigenicity.* — Genetics, v.125, No.3, pp.633–644.

Kimmel M. and Axelrod D.E. (1994): *Fluctuation test for two-stage mutations: application to gene amplification.* — Mutat. Res., v.306, No.1, pp.45–60.

Kimmel M., Axelrod D.E. and Wahl G.M. (1992): *A branching process model of gene amplification following chromosome breakage.* — Mutat. Res., v.276, No.2, pp.225–240.

Kimmel M. and Stivers D.N. (1994): *Time-continuous branching walk models of unstable gene amplification.* — Bull. Math. Biol., v.56, No.2, pp.337–357.

Luria S.E. and Delbrück M. (1943): *Mutations of bacteria from virus sensitivity to virus resistance.* — Genetics, v.28, pp.491–511.

Marx J. (1993): *New colon cancer gene discovered.* — Science, v.260, pp.751–752.

Richards R.I. and Sutherland G.R (1992): *Dynamic mutations: a new class of mutations causing human disease.* — Cell, v.70, pp.709–712.

Searls D.B. (1992): *The linguistics of DNA.* — Amer. Scient., v.80, No.6, pp.579–591.

Smith G.P. (1976): *Evolution of repeated DNA sequences by unequal crossover.* — Science, v.191, pp.528–535.

Smith K.A., Stark M.B., Gorman P.A. and Stark G.R (1992): *Fusions near telomeres occur very early in the amplification of CAD genes in Syrian hamster cells.* — Proc. Natl. Acad. Sci. USA, v.89, pp.5427–5431.

Stark G.R. (1993): *Regulation and mechanisms of mammalian gene amplification.* — Adv. Cancer Res., v.61, pp.87–113.

Stephan W. (1989) *Tandem-repetitive noncoding DNA: forms and forces.* — Molecular Biology and Evolution, v.6, pp.198–212.

Tlsty T., Margolin B.H. and Lum K. (1989): *Differences in the rates of gene amplification in nontumourigenic and tumourigenic cell lines as measured by Luria-Delbrück fluctuation analysis.* — Proc. Natl. Acad. Sci. USA, v.86, pp.9441–9445.

Windle B, Draper B.W., Yin Y., O'Gorman S. and Wahl G.M. (1991): *A central role for chromosome breakage in gene amplification, deletion, formation, and amplicon integration.* — Gene Dev., v.5, pp.160–174.

Windle B. and Wahl G.M. (1992): *Molecular dissection of mammalian gene amplification: New mechanistic insights revealed by analysis of very early events.* — Mutat. Res., v.276, pp.199–224.

Wheldon T.E. (1988) *Mathematical Models in Cancer Research.* — Philadelphia: Adam Hilger.