LA-UR 91 - 4129

LA-UR--91-4129

DE92 004858

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36

TITLE ROLES OF REPETITIVE SEQUENCES

AUTHOR(S) GEORGE I. BELL

SUBMITTED TO : PROCEEDINGS OF THE WORKSHOP ON OPEN PROBLEMS IN
COMPUTATIONAL MOLECULAR BIOLOGY, TELLURIDE, CO,
JUNE 2-8, 1991, A. KONOPKA, EDITOR

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes.

The Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy.

Los Alamos Los Alamos National Laboratory
Los Alamos, New Mexico 87545

LAUR-91-

ROLES OF REPETITIVE SEQUENCES*

by

George I. Bell

Theoretical Biology and Biophysics (T-10)

Theoretical Division

Los Alamos National Laboratory

Los Alamos, NM 87545 U.S.A.

a Talk Presented at the

Workshop on Open Problems in Computational Molecular Biology

Telluride, CO, June 2 - 8, 1991

T_EX:Reutenauer, rs

* Work supported by the U. S. Department of Energy.

ROLES OF REPETITIVE SEQUENCES

George I. Bell

Introduction

The DNA of higher eukaryotes contains many repetitive sequences. The amount of such DNA in a genome and information about copy number can be estimated from measurements of the reassociation kinetics – the Cot curve – of genomic DNA (Lewin, 1990, Ch. 24). In mammalian genomes, roughly half of the DNA may be regarded as repetitive. Genome projects will naturally focus on the discovery of genes, which are for the most part single copy sequences. The exceptions to this rule are of two kinds. First, there are a few genes, such as those for ribosomal RNA and histones, that are present in multiple copies because of the large amounts of gene product that the cell requires. Second, many and perhaps most genes are members of multigene families, where the members of a family may have evolved from a common ancestor by such processes as gene duplication, mutation, recombination, genetic drift, and selection. The sequences of such family members have usually diverged enough that they will show up as single-copy DNA in reassociation kinetics, and they will be regarded as single copy in this paper. The study of such families may be expected to provide deep insights into the processes of genome evolution. However, the study of repetitive sequences will also be important, not only because many have important biological function, but also because they provide information on genome organization, evolution and dynamics.

Repetitive DNA can be classified in many ways; for example, by copy number which may range from several to millions. Or by whether the repeats are tandemly

arrayed or interspersed, and if in a tandem array, where they are located on a chromosome. The length of the repeat, whether it is simple (short) or complex (long), and the sequence and degree of sequence conservation among related members of a repeat family are also important and may provide information regarding the origin and evolution of the family. Finally, we would like to know whether the repeats perform some function for the cell or if they are examples of selfish or parasitic DNA (Doolittle and Sapienza, 1980; Orgel and Crick, 1980), concerned with their own propagation and survival.

In this paper, I will first discuss some generic effects that repetitive sequences will have upon genome dynamics and evolution. In particular, it will be shown that repetitive sequences foster recombination among, and turnover of, the elements of a genome. I will then consider some examples of repetitive sequences, notably minisatellite sequences and telomere sequences as examples of tandem repeats, without and with respectively known function, and Alu sequences as an example of interspersed repeats. Some other examples will also be considered in less detail. In the appendix a simple theoretical birth and death model is shown to produce homogenization among the members of a tandem multigene family, and results are given for the probability that two family members are identical by descent, and applied to a particular minisatellite locus.

Repetitive Elements Foster Genome Recombination and Turnover

The processes that make this happen have been collectively called "molecular drive" by G. Dover (1982, 1986). A number are known to exist though the molecular

basis of none, in eukaryotes, is well understood. Processes to be considered include crossing over, gene conversion, and transposition.

Crossing over between homologous chromosomes during meiosis is a common occurrence. In human chromosomes, two markers separated by about 10^6 bases will, on average, have a one percent chance of a crossing over between them each meiosis. Since the human haploid genome contains 3×10^9 bases, we thus have about 30 crossovers per meiosis. Crossing over also takes place between sister chromatids during mitosis, perhaps at about ten percent of the rate during meiosis (Wolf and Perry, 1974), and if the mitoses occur in the germline, the results can be passed on to further generations.

Let us consider an array of tandemly repeated sequences on a chromosome. As the chromosomes become paired prior to a potential crossover in the repeat region, it is likely that they will have difficulty in establishing exactly the right alignment; misspairing by one or more repeats may be expected to be almost as likely as exact pairing. The result of a crossover misspaired by X repeats will be one chromosome (or chromatid in mitosis) that is longer by X repeat units and one that is shorter by X repeats, as shown in Fig. 1. Hence, unequal crossing over will lead to a dispersion in the number of repeats. Moreover, if not all the repeats are the same (due to mutations), this random process will tend to lead to the elimination of most of the members of the repeat family and retention of only one or a few versions in multiple copies. The result is "concerted evolution," the remarkable tendency for family sequence conservation within a species, yet divergence between even closely related species. Some examples may be found in tandemly arrayed

genes for ribosomal RNA and histones, where the sequence conservation extends to "non-functional" spacer regions as well as the functional DNA.

The process of gene fixation by repeated rounds of unequal crossing over was first analyzed in computer Monte Carlo simulation (Smith, 1973; Black and Gibson, 1974). Smith, 1976, also pointed out that any sequence that can undergo unequal crossover will tend to develop periodic base patterns, i.e., repeats. An analytical treatment was developed by Ohta, 1976, using a diffusion equation approach from population genetics, and considering a single chromosome lineage with unequal mitotic crossover by one repeat, one gain, and one loss per unit time. She found the fixation time to be the square of the number of repeats in the family where the total number of repeats was assumed to remain constant. (For misspairing by up to ℓ repeats, Ohta estimated the fixation time to be reduced by roughly ℓ .) Similar results were obtained by Perelson and Bell, 1977, using a birth and death model for repeat evolution. This model is summarized in Appendix A; missprints in the original article rendered some portions unintelligible. In a more realistic model, in which members of the repeat family are subject to mutation, to inter-chromosomal equal and unequal crossover at meiosis, unequal intra-chromosomal crossover at mitosis and random genetic drift, Ohta, 1980, was able to obtain some partial results using diffusion methods from population genetics. In particular, she was able to characterize the equilibrium identity coefficient or probability that two members of the family are identical by descent (see Appendix B). These results were applied to several systems, including immunoglobulin genes (Ohta, 1980) and

to the genes for ribosomal RNA and histones in *Drosophila* (Coen et al., 1982). For a brief review of the genes for human rRNA, see Dover, 1989.

Interspersed repeats can also lead to unequal crossing over and duplication or elimination of the DNA between the repeats. Many instances of genetic disease in humans have been found to arise from unequal crossovers between Alu repeats in the affected genes. For example, a case of familial hypercholesterolemia was found to result from recombination between an Alu sequence in intron 15 with an Alu sequence in the 3' untranslated region of the last exon of the LDL receptor gene (Lehrman et al., 1987). A number of other instances are cited in Lehrman et al., 1987. Recombination between Alu, or other repeat, sequences flanking a gene can give rise to gene duplication of possible importance for the evolution of multigene families (e.g., Chen et al., 1989, discuss the possible role of Alu repeats in evolution of the five genes of the human growth hormone locus.).

A different type of recombination event is "gene conversion." Although the molecular basis of gene conversion is unknown, the result is that after meiosis, two alleles are unequally expressed in the haploid products, even though they were each present in the homologous chromosomes of the pre-meiotic cell (Lewin, 1990, Ch. 32). These results are usually interpreted in terms of a breakage of one or both strands of DNA on one of the chromosomes, followed by a misspairing with the other chromosome to form a heteroduplex, which is then resolved by "correcting" one of the strands. As shown in Fig. 2, depending on how the intermediate four strand "Holliday junction" is cut, one can form a heteroduplex with or without recombination of the distal DNA portions. During resolution of the heteroduplex,

perhaps by DNA repair enzymes, it is easy to imagine that one or the other strand will be favored so that the gene conversion will be biased in favor of one allele. Insofar as gene conversion may operate between genes on different chromosomes, e.g., histone genes in animals, one may speculate that an RNA intermediate might be involved as an alternative to (miss)pairing of non-homologous chromosomes. However, no evidence for an RNA intermediate has been found.

Transposition of pieces of DNA is widespread in the genomes of cells, prokaryotic, and eukaryotic (Lewin, 1990, Chs. 33 and 34). Transposition events can be conservative or replicative depending on whether or not there is an increase in the number of copies of the transposed DNA. During replicative transposition, the number of copies is increased, i.e., a repeat family is being created or expanded. Both conservative and replicative transposition may or may not involve an RNA intermediate. When transposition involves an RNA intermediate, we refer to the product as a retroposon. The multiplication of a retroposon has many similarities to the life cycle of a retrovirus; RNA transcripts, made from DNA in the usual way, are reverse-transcribed to DNA and reinserted in the genome. Evidently the multiplication of retroposons must be limited in a cell in order to maintain cell viability.

The processes that we have been discussing may also be classified as either random or directional. Random processes would include equal and unequal crossing over between homologous chromosomes at meiosis; unequal crossing over between sister chromatids at mitosis, unbiased gene conversion, and conservative transposition. In these random processes, there is no intrinsic tendency to increase/decrease

the number of members of a repeat family or to favor one allele over another. In contrast, directional processes include gene duplication, biased gene conversion, and replicative transposition. Of course, random processes when combined with selection can lead to non-random results. For example, unequal crossing over between two Alu repeats flanking an essential gene may be equally likely to lead to gene loss or to gene duplication. Selection, however, will favor survival of germ cells with the gene duplication as opposed to the gene loss so that duplicated genes will appear to result from random unequal crossover.

Although the above mechanisms have provided plausible explanations of concerted evolution, clear and quantitative data concerning any one of them have been hard to come by, in part because of a lack of molecular understanding. A recent report (Hillis et al., 1991) provides evidence for rapid biased gene conversion in concerted evolution of ribosomal DNA genes, for some species of parthenogenetic lizards. Whether such results hold for other species or genes is unknown. However, it is to be expected that as further sequence information becomes available from genome projects, the study of repetitive sequences will reveal much more concerning mechanisms of molecular drive and genome evolution.

Some Simple Tandem Repetitive Sequences

These are tandem arrays of relatively simple repeats, sometimes present on many chromosomes. The functions, if any, of most are unknown, but in the case of telomeric DNA there is a clear function. The following selection is meant to be illustrative rather than comprehensive.

Satellite Sequences are multiple repeats of short sequences that, because of their abundance and uniformity of sequence, show up as satellite peaks when genomic DNA is centrifuged in a density gradient of CsCl. They are typically associated with centromeric regions of eukaryotic chromosomes, vary greatly in amount and sequence between different species, and have no known function. As an extreme example, in *Drosophila virilis*, one 7-base repeat in 10^7 copies makes up 25% of the genome, while two others differing by one base each, make up an additional 15% (Lewin, 1990, ch. 28). Mammalian satellite DNA has a hierarchy of repeats. For example, much of mouse satellite DNA has a 234 base period, which, however, seems to be built from a 9 base repeat that has undergone duplication and mutation many times. Evidently, satellite DNA could expand or contract by unequal crossing over, but the system is difficult to sequence or study because of the long stretches of repetitive DNA that are involved. It has been argued by Smith, 1976, that satellite DNA probably arose by unequal crossing over and a model based on sister chromatid exchange was examined in some detail. It may be noted that Smith assumed that the time to fixation was proportional to n , the number of repeat copies, but this assumption is not inconsistent with the estimates of Ohta, 1976, when the extent of misspairing is also assumed proportional to n . Of more concern, similar satellite DNA is found near the centromeres of many chromosomes, suggesting that a mechanism for interchromosomal exchange of centromeric material is required. In addition, the rate of crossover is believed to be much reduced in centromeric regions. Thus it appears that the evolution of satellite DNA is still not well understood.

Minisatellite Sequences are tandem repeats that are much less extensive than satellites. There are many families interspersed in the genome, and because at a given locus the number of repeats is usually highly polymorphic, they have proved to be excellent markers for genetic mapping based on restriction fragment length polymorphism. Although the length of an allele at any locus is usually inherited in stable Mendelian fashion, occasional "mutations" to new length alleles can be traced in family studies (Jeffreys et al., 1988). For highly variable loci, extremely high mutation rates (up to $\sim 10\%$ per gamete for one locus) are observed, consistent with a very high rate of unequal crossover within the locus or some form of "slippage" during DNA replication. For less variable loci, the mutation rates were significantly less. Comparable mutation rates were observed for sperm and oocytes, suggesting that the mutations may be taking place during meiosis since the oocyte and sperm lineages would have very different numbers of postzygotic mitoses. However, in more recent studies of a particular locus, in which the individual repeats have a highly informative pattern of variation (i.e., a fraction contain a Hae III restriction site, while the rest do not), Jeffreys et al., 1990, were able to analyze the parentage of individual length mutants found in the sperm and blood cells of an individual heterozygous at this locus. In no case was there evidence of a crossover between homologous parental chromosomes. On the contrary, all the mutants were consistent with unequal sister chromatid exchange or some other process that changes the number of repeats on a chromosome while retaining the flanking markers. Thus the alleles at this locus appear to evolve along haploid chromosomal lineages; lack of interchromosomal crossover will thus frustrate concerted

evolution. The mutation rate at this locus (~ 4.5 kb long, in the individual studied) was $\sim 1.4 \times 10^{-3}$ per sperm.

It is of interest to compare this mutation rate with the rate of mitotic unequal crossover that might be expected at this locus. If we assume that the mitotic crossover rate is about one-tenth the average meiotic rate, thus $\sim 10^{-9}$ /bp, and that most crossovers in a long repetitive locus will be unequal, then the rate of mitotic unequal crossover for this locus would be $4.5 \times 10^3 \text{ bp} \times 10^{-9} = 4.5 \times 10^{-6}$ per mitosis. It has also been estimated that the average mature sperm has undergone 400 post zygotic mitoses (Vogel and Rathenberg, 1975). Hence, the expected probability of a length mutant is about $400 \times 4.5 \times 10^{-6} = 1.8 \times 10^{-3}$, in very reasonable agreement with the result of Jeffreys et al., 1990. Independent studies of the same locus (Armour et al., 1989; Jeffreys et al., 1991) have indicated that in tumor cells, the unequal crossover rate may be $\sim 4 \times 10^{-6}$ per mitosis, while paternity studies give the probability of new length mutants $\sim 1.2 \times 10^{-2}$ per sperm or oocyte. While the latter number is higher than our estimate above, it would not be surprising if recombination were enhanced above the average at a long repetitive locus, inasmuch as the opportunities for strand pairing are more numerous than for single copy DNA.

Is this rate of unequal crossover sufficient to maintain sequence homogeneity in the repeat family in the face of point mutations? This question is considered in Appendix B where it is estimated that the probability that two repeats are identical by descent is roughly 0.6. This is in good agreement with the observation of two major repeat alleles, with or without the Hae III site.

It thus appears that while a consistent pattern of minisatellite evolution has yet to appear, and indeed the importance of various mechanisms may vary with the locus, nevertheless the tools are at hand for detailed studies of mutational patterns and comparison with theoretical models.

Telomeric Repeats are found at the ends of eukaryotic chromosomes and have an important function in preserving chromosome length. In the absence of special ending sequences, structures, and/or mechanisms, chromosomes would shorten with each round of DNA replication (since DNA synthesis cannot replicate the 5' end of a strand) and would also be exposed to the action of exonucleases. The basic telomeric repeat is TTAGGG in mammals (Moyzis et al., 1988; Moyzis, 1991), and there are similar G-rich sequences in other species on the strand with a 3' end. On human chromosomes, there are between 250 and 1500 repeats per telomere. These G-rich telomeric sequences form special structures, probably overhanging to fold back from the ends and, when two chromatids are present, possibly forming four stranded structures with Hoogsteen base pairing between the guanines (Sundquist and Klug, 1988; Sen and Gilbert, 1990; and A. Rich, unpublished crystallographic data). These structures presumably protect the chromosomes from exonucleases. Moreover, telomers are apparently recognized by a special reverse transcriptase, telomerase, which adds one or more copies of the repeat sequence to the 3' end of the DNA, using an RNA template with the complementary sequence (Shippen-Leutz and Blackburn, 1990).

Telomeric repeats are clearly a special case, but their examination indicates that some repeats may form special structures of functional significance. In addition, we

observe that cells contain reverse transcriptases other than those associated with retroviruses, a point of some interest when we consider some retroposons.

(GT)_n is a simple repeat, widely interspersed in the genomes of all eukaryotes. At a given locus, the length, *n*, is highly polymorphic (Weber and May, 1989), as for many minisatellite sequences of which the GT repeat may be considered an especially simple member. There are estimated to be around 100,000 (GT)_n loci in mammals (Stallings et al., 1991). These repeats are of considerable interest, apart from their use as polymorphic markers, because of their potential to form left-handed or ZDNA under negative supercoiling of the DNA (Nordheim and Rich, 1983). It has been suggested that they may play a role in determining chromosome structure or in genetic recombination (Pardue et al., 1987) or in the regulation of gene transcription (Hamada et al., 1984), but their functional significance remains controversial. Stallings et al., 1991, have examined the extent to which (GT)_n loci are conserved between species, as deduced from genomic sequences in GenBank. They conclude that although there is strong conservation of position between man and apes or between rat and mouse, the lesser correlation (~ 30%) between man and rodent indicates that GT repeats do not play a critical role in regulation of gene expression. If length polymorphism at a GT locus is maintained by unequal crossover within the locus, then a high rate of crossover would help to maintain repeat fidelity against point mutations, as noted for minisatellite sequences.

In this context, it may be noted that Waldman and Liskay, 1988, claim that a long uninterrupted stretch (~ 200bp) of nearly perfect sequence homology is required for efficient recombination in mammalian cells. This result would suggest

that GT repeats are unimportant in recombination since they are generally short compared to 200 bp. However, it is possible that the Waldman and Liskay result is not a general one – the length for recombination recognition is ~ 30 bp in *E. coli* – or that $(GT)_n$ forms recombination-prone structures (Pardue et al., 1987). Moreover, as noted in the following section, unequal crossovers between Alu sequences with numerous sequence differences appear to be common.

CCG repeats have been localized to the fragile X site, the most common locus associated with mental retardation (Kremer et al., 1991; Annemieke et al., 1991). This repeat appears to be in a gene exon and to have ~ 30 copies in normal individuals but substantially more in affected individuals. Instability of length is observed *in vitro* and the sequence is difficult to work with, probably because of the very high GC content. Kremer et al., 1991, speculate that the repeat may have difficulty replicating *in vivo*. This site is the current subject of intense study, and we may expect much more to be reported about the peculiar CCG repeat in the near future.

Alus and Other Retroposons

Interspersed repetitive elements are ubiquitous in eukaryote genomes. Many show signs of having arisen via replicative transposition using an RNA intermediate; i.e., they are retroposons. In the human genome, the most abundant of these interspersed elements are the Alu sequences, so-called because most are cleaved by the restriction nuclease, Alu I. There are between 5×10^5 and 10^6 copies of the Alu sequence in the human haploid genome (Hwu et al., 1986; Moyzis et al., 1989); some genes contain up to 25% of their DNA in the form of Alu sequences (e.g., Chen et al., 1989) in their introns and 5' or 3' untranslated regions. Alu sequences comprise

between 5% and 10% of human DNA. A typical member, Fig. 3, is around 300 bp in length, with two tandemly arranged 130 bp portions, with the right one containing an extra 31 bp insertion. There is considerable ($\sim 15\%$) sequence variability between members of the family, but most contain a direct repeat, of variable sequence and length, at the two ends, which is taken as evidence of an insertion event via a staggered nick in host DNA. Also, usually present is an A rich or poly A tail to each of the subunits, suggestive of an RNA intermediate. Many have an RNA Polymerase III transcription site - recall that tRNA and other small RNA genes are transcribed by Polymerase III and contain an internal promoter.

By sequence comparison, it has been deduced that Alu sequences have been derived from the 7sL RNA gene (Weiner, 1980; Ulla and Tschu $\ddot{u}$ g, 1984). That gene product is an abundant cytoplasmic RNA that, together with various proteins, mediates the transport of secretory proteins across the endoplasmic reticulum. Other short interspersed elements (SINEs) have been shown to be similar in sequence to transfer RNA genes, especially lysine tRNA, and it has been argued that Alus, other SINEs, and tRNAs have many similarities of secondary structure (see Okada, 1990). In this connection, it is of interest to recall that viral reverse transcriptases use host tRNA as a primer.

Analysis of the sequences of various Alus has made it possible to classify them into various families, including one family that, because of the homogeneity of its members, appears to be of relatively recent origin (e.g., Batzer et al., 1990; Britten et al., 1989; Batzer and Deininger, 1991).

Since many Alu sequences are found in the introns or untranslated ends of genes, there are in a human cell nucleus abundant RNA transcripts containing Alus. Few, however, survive RNA processing to emerge in mature mRNA. A plausible hypothesis for the many Alu sequences in the DNA of humans (and apes) is that during their expansion phase at least some were being transcribed by RNA Polymerase III, using the internal promoter, and that the transcripts were reverse transcribed by some, probably unknown, enzyme and reinserted more or less at random into the genome. Is there evidence of such activity at present? A few examples of polymorphic insertions of Alus have been reported (see Matera et al., 1990). Moreover, some polyadenylated RNAs have been found that resemble Alu transcripts (Matera et al., 1990). It thus appears that a few of the Alu sequences may be active and able to get themselves transcribed and inserted, but that the great majority are inactive processed pseudogenes.

Various functions have been suggested for Alu sequences, including that they serve as origins of replication, but none has been confirmed by experiment. Of course, as noted earlier, the mere presence of interspersed Alus fosters unequal recombination with deletion or duplication of genetic material. Such turnover may be a significant evolutionary process.

It is of some interest to try to estimate the frequency of unequal recombination involving Alu elements. Suppose that Alus represent 10% of human DNA and that 10% of the recombination events take place at Alu sites. Thus there would be about three Alu recombinations per meiosis and 0.3 per mitosis. What fraction might be unequal? Suppose the initial mispairing of chromosomes or chromatids is $\sim 1/10$

kb. (Recombination between Alus separated by ~ 10 kb have been reported, e.g., by Lehrman et al., 1987.) Within ± 10 kb of a given Alu there are likely to be about six more. Of these, about half will have the opposite orientation; if such recombination takes place, it would produce grossly aberrant chromosomes and presumably unviable cells. We ignore such recombinants. Of the remaining three Alus with correct orientation, often the sequence homology between Alus may be insufficient to allow efficient recombination. Suppose, on average, a misspaired Alu has 10% of the recombination rate of an exact copy. Then we would expect the fraction of Alu crossovers that are unequal to be $\sim 0.3/1.3 \simeq 0.2$. Thus we estimate, very roughly, 0.6 unequal crossovers per meiosis and 0.06 per mitosis. For 400 postzygotic mitoses per sperm (Vogel and Ruthenberg, 1978), this would give 24 unequal Alu crossovers due to mitoses in the germ line per sperm. While these estimates are clearly very uncertain, they do point to possibly high rates of genome turnover due to Alu recombination.

Another important family of interspersed sequences in human DNA is the L1 (or Kpn 1) sequence. This is an example of a long interspersed sequence (LINE); a full length member is about 6.5 kb in length, though most have lost some of their 5' ends. There are between 50,000 and 100,000 copies in the human genome (Hwu et al., 1986; Moyzis et al., 1989). There are similarities between L1 and retroviral sequences, suggesting that L1 may code for a reverse transcriptase. Some L1 sequences are transcriptionally and transpositionally active; insertion of L1 sequences has been responsible for human genetic disease.

Interspersed repetitive elements are found in all eukaryotes. Among the best studied are Ty elements in yeast. These are about 6 kb in length and present in around 40 cy in a yeast genome (Lewin, 1990, Ch. 34). Some appear to code for a reverse transcriptase; they are transcribed and transposed. A somewhat similar family in *Drosophila* is called Copia. Among other families in *Drosophila* are the P elements, up to 9 kb in length and present in up to 50 copies per genome. They appear to transpose without an RNA intermediate and to code for a DNA transposase. P elements may play some role in speciation in *Drosophila*; crosses between males having P elements and females without them are sterile. In all the above systems, the presence of repeats fosters unequal recombination and genome turnover.

Functional Multigene Families

Various genes are present in multiple copies in eukaryotic DNA, notably those for ribosomal RNA and histones. Presumably the need for large quantities of gene product has produced a requirement for multiple gene copies. In a more complete review of repetitive sequences, such families would be considered in detail, and the arguments would be given for how molecular drive maintains sequence homogeneity, within a species in such families, not only of the sequences that code for useful gene products but also, to a considerable extent, of the spacer regions between them. However, in the limited time for this talk, I merely note these important families and refer to some original papers for further discussion. See Ohta, 1980, for discussion of immunoglobulin genes and Coen et al., 1982, for discussion of ribosomal RNA and histone genes in *Drosophila*.

Concluding Remarks

The foregoing catalog of repetitive sequences is illustrative only, and far from complete. Repetitive sequences, both tandem and interspersed, are ubiquitous in eukaryote DNA. Their presence provides evidence of mechanisms that operate during the evolution of genomic DNA, and, once established, repetitive elements promote the turnover of genomic DNA. By this means, repetitive elements play important roles in the evolution of genomes.

Acknowledgement

I am indebted to Yves Quentin for acquainting me with some of the recent literature on Alu sequences.

APPENDIX A

Birth and Death Model for Evolution of Tandem Repeats

This material is taken from Perelson and Bell, 1977, with some simplifications and corrections of numerous misprints. Consider a family of N repeats on a chromosome. For simplicity, consider only sister chromatid exchange with the possibility of misspairing by one repeat, leading, as shown in Fig. 1, to gain or loss of one repeat. Let us assume that not all the repeats are the same, say because of mutation, but that they occur in subfamilies ($i = 1, 2, \dots, I$), where all members of the subfamily are identical and the i^{th} subfamily is represented n_i times. Confining attention to this haploid lineage, i.e., a series of mitoses in the germ lines of successive generations, let us define the unit of time such that the probability of a crossing over with unit misspairing is, in the time δt , equal to δt . Let the probability of unit increase or decrease in n_i be $n_i/2N\delta t$. (I.e., the probability that the crossover affects the i^{th} subfamily is equal to its proportional representation, n_i/N , in the family, and it is equally likely to gain or lose a member.) If $P_{n_i}(t)$ represents the probability that n_i copies are present at time t , we have

$$\frac{dP_{n_i}(t)}{dt} = -\frac{n_i}{N}P_{n_i}(t) + \frac{n_i-1}{2N}P_{n_i-1}(t) + \frac{n_i+1}{2N}P_{n_i+1}(t) \quad ; \quad n_i = 1, 2, \dots, N \quad . \quad (A1)$$

and

$$\frac{dP_0(t)}{dt} = \frac{1}{N}P_1(t) \quad . \quad (A2)$$

If initially there is one copy of each subfamily present, the initial conditions are $P_{n_i}(0) = 1$ if $n_i = 1$, $= 0$ if $n_i \neq 1$.

These equations define a linear birth and death process with known solution (e.g., Feller, 1968),

$$\begin{aligned} P_0(t) &= t/(t + 2N) \\ P_{n_i}(t) &= \frac{t/2N)^{n_i-1}}{(1 + t/2N)^{n_i+1}} \quad ; \quad n_i \geq 1 \quad . \end{aligned} \quad (43)$$

Note that at very long times, each subfamily will become extinct, i.e.,

$$\lim_{t \rightarrow \infty} P_0(t) = 1 \quad .$$

However, it was argued that before this happens, there will be a time during which one subfamily remains, which may be called gene fixation. The mean time to gene fixation was estimated as the time when the mean number of extinct genes (subfamilies) was $N - 1$, i.e., $N P_0(t) = N - 1$. This gives a fixation time of $2N(N - 1)$.

Ohta, 1976, obtained similar results for a diffusion model of gene fixation using the models of population dynamics. Note that her elementary event involved two crossovers, one involving a gain of one repeat and the other a loss, so that her time was N^2 , in agreement with ours for large N . Ohta also argued that if the crossover mismatch involves ℓ repeats on average, then the fixation time should be reduced by ℓ . The above results were in rough agreement with numerical simulations of Smith, 1974, and Black and Gibson, 1974.

It appears straightforward, if complicated, to generalize the birth and death formulation to more realistic models of family evolution, but, to my knowledge, no solutions have been obtained.

APPENDIX B

Probability of Identity by Descent

Consider the previous model of n tandem repeats subject to unequal crossing over but now also subject to mutation as well. Assume that each mutation gives rise to a new subfamily; i.e., the number of possible mutations is much larger than the number of subfamilies. Let X_i be the probability that a repeat belongs to subfamily i (i.e., $X_i = n_i/n$). The probability C , that two genes are identical by descent, is then

$$C = \sum_{i=1}^{\infty} X_i^2 . \quad (B1)$$

(C is sometimes also called the clonality.) As the repeat family evolves, some subfamilies are becoming extinct while others are being generated by mutation. At steady state, Ohta, 1980, estimated C for rather general crossing over models. In particular, for the haploid lineage, if v = mutation rate per repeat, γ = unequal mitotic crossover rate per family, and m = mean number of repeats misspaired, then at steady state

$$C = \frac{m\gamma}{2vn^2 + m\gamma} . \quad (B2)$$

It is of interest to estimate C for the minisattelite locus of Jeffreys et al., 1990. We have earlier estimated $\gamma \simeq 4.5 \times 10^{-6}$ per mitosis, and if there are an average of around 10 mitoses per year in the sperm lineage, we have $\gamma \simeq 5 \times 10^{-5}$ per year. The mutation rate per base may be estimated as 1.5×10^{-9} per year (Britten, 1986), so that for a repeat of 29 bp, $v \simeq 4.5 \times 10^{-8}$ per year. For the individual considered $n \simeq 150$, and it appeared that $m \simeq 50$. Hence, for this locus we estimate

$C \simeq 0.6$. This seems a very reasonable result since for this locus two major repeat alleles have been identified; two equally likely alleles would give $C = 0.5$.

REFERENCES

- Annemieke, J.M.H., et al., 1991. Identification of a Gene (FMR-1) Containing a CGG Repeat Coincident with a Breakpoint Cluster Region Exhibiting Length Variation in Fragile X Syndrome. *Cell* 65:905-914.
- Armour, J.A.L, I. Patel, S.L. Thein, M.F. Fey, and A.J. Jeffreys, 1989. Analysis of Somatic Mutations at Human Minisatellite Loci in Tumors and Cell Lines. *Genomics* 4:328-334.
- Black, L.A. and D. Gibson, 1974. Neutral Evolution and Immunoglobulin Diversity. *Nature* 250:327-328.
- Britten, R.J., 1986. Rates of DNA Sequence Evolution Differ Between Taxonomic Groups. *Science* 231:1393-1398.
- Britten, R.J., D.B. Stout, and E.R. Davidson, 1989. The Current Source of Human Alu Retroposons Is a Conserved Gene Shared with Old World Monkeys. *PNAS USA* 86:3718-3722.
- Chen, E.Y., Y.-C. Liao, D.H. Smith, H.A. Barrera-Saldaña, R.E. Gelinas, and P.H. Seeburg, 1989. The Human Growth Hormone Locus: Nucleotide Sequence, Biology, and Evolution. *Genomics* 4:479-497.
- Coen, E., T. Strachan, and G. Dover, 1982. Dynamics of Concerted Evolution of Ribosomal DNA and Histone Gene Families in the *melanogaster* Species Subgroup of *Drosophila*. *J. Mol. Biol.* 158:17-35.
- Doolittle, W.F. and C. Sapienza, 1980. Selfish Genes, the Phenotypic Paradigm, and Genome Evolution. *Nature* 284:601-603.
- Dover, G.A., 1982. Molecular Drive: A Cohesive Mode of Species Evolution. *Nature* 299:111-117.
- Dover, G., 1986. Molecular Drive in Multigene Families: How Biological Novelty Arise, Spread, and Are Assimilated. *TIG*, June, 159-165.
- Dover, G.A., 1989. Linkage Disequilibrium and Molecular Drive in the rDNA Gene Family. *Genetics* 122:249-252.
- Hamada, H., M. Seidman, B.H. Howard, and C.M. Gorman, 1984. Enhanced Gene Expression by the Poly (dT-dG) · Poly(dC-dA) Sequence. *Mol. & Cell Biol.* 4:2622-2630.
- Hillis, D.M., C. Moritz, C.A. Porter, and R.J. Baher, 1991. Evidence for Biased Gene Conversion in Concerted Evolution of Ribosomal DNA. *Science* 251:308-310.
- Hwu, H.R., J.W. Roberts, E.H. Davidson, and R. Britten, 1986. Insertion and/or Deletion of Many Repeated DNA Sequences in Human and Higher Ape Evolution. *PNAS USA* 83:3875-3879.
- Jeffreys, A.J., N.J. Royle, V. Wilson, and Z. Wong, 1988. Spontaneous Mutation Rates to New Length Alleles at Tandem-Repetitive Hypervariable Loci in Human DNA. *Nature* 332:278-281.
- Jeffreys, A.J., R. Neuman, and V. Wilson, 1990. Repeat Unit Sequence Variation in Minisatellites: A Novel Source of DNA Polymorphism for Studying Variation and Mutation by Single Molecule Analysis. *Cell* 60:473-485.

- Jeffreys, A.J., A. MacLeod, K. Tamaki, D.L. Neil, and D.G. Monckton, 1991. Minisatellite Repeat Coding As a Digital Approach to DNA Typing. *Nature* 354:204-209.
- Kremer, E.J. et al., 1991. Mapping of DNA Instability at the Fragile X to a Trinucleotide Repeat Sequence p(CCG)_n. *Science* 252:1711-1714.
- Lehrman, M.A., D.W. Russel, J.L. Goldstein, and M.S. Brown, 1987. Alu-Alu Recombination Deletes Acceptor Sites and Produces Secreted Low Density Lipoprotein Receptor in a Subject with Familial Hypercholesterolemia. *J. Biol. Chem.* 262:3354-3361.
- Lewin, B. 1990. *Genes IV*. Oxford: Oxford University Press.
- Matera, A.G., U. Hellmann, and C.W. Schmid, 1990. A Transpositionally and Transcriptionally Competent Alu Subfamily. *Mol. & Cell Biol.* 10:5424-5432.
- Matzer, M.A., G.E. Kilroy, P.E. Richard, T.H. Shaikh, T.D. Desselle, C.L. Hopkins, and P.L. Deininger, 1990. Structure and Variability of Recently Inserted Alu Family Members. *Nucl. Acids Res.* 18:6793-6798.
- Moyzis, R.K., J.M. Buckingham, L.S. Cram, M. Dani, L.L. Deaven, M.D. Jones, J. Meyne, R.L. Ratliff, and J-R. Wu, 1988. A Highly Conserved Repetitive DNA Sequence, (TTAGGG)_n, Present at the Telomeres of Human Chromosomes. *PNAS USA* 85:6622-6626.
- Moyzis, R.K., D.C. Torney, J. Meyne, J.M. Buckingham, J-R. Wu, C. Burks, K.M. Sirotkin, and W.B. Goad, 1989. The Distribution of Interspersed Repetitive DNA Sequences in the Human Genome. *Genomics* 4:273-289.
- Moyzis, R.K., 1991. The Human Telomere. *Sci. Am.* August:48-55.
- Nordheim, A. and A. Rich, 1983. The Sequence (dC-dA)_n·(dT-dG)_n Forms Left-Handed Z-DNA in Negatively Supercoiled Plasmids. *PNAS USA* 80:1821-1825.
- Ohta, T., 1976. Simple Model for Treating Evolution of Multigene Families. *Nature* 263:74-75.
- Ohta, T., 1980. *Evolution and Variation of Multigene Families*. Berlin: Springer-Verlag.
- Okada, N., 1990. Transfer RNA-Like Structure of the Human Alu Family: Implications of Its Generation Mechanism and Possible Functions. *J. Mol. Evol.* 31:500-510.
- Orgel, L.E. and F.H.C. Crick, 1980. Selfish DNA: The Ultimate Parasite. *Nature* 284:604-607.
- Pardue, M.L., K. Lowenhaupt, A. Rich, and A. Nordheim, 1987. (dC-dA)_n·(dG-dT)_n Sequences Have Evolutionarily Conserved Chromosomal Locations in *Drosophila* with Implications for Roles in Chromosomal Structure and Function. *EMBO J.* 6:1781-1789.
- Perelson, A.S. and G.I. Bell, 1977. Mathematical Models for the Evolution of Multigene Families by Unequal Crossing Over. *Nature* 265:304-310.
- Sen, D. and W. Gilbert, 1990. A Sodium Potassium Switch in the Formation of Four-Stranded G4-DNA. *Nature* 344:410-414.
- Shippen-Lentz, D. and E. Blackburn, 1990. Functional Evidence for an RNA Template in Telomerase. *Science* 247:546-552.

- Smith, G.P., 1974. Unequal Crossover and the Evolution of Multigene Families. *Cold Spring Harbor Quart. Biol.* 38:507-514.
- Smith, G.P., 1976. Evolution of Repeated DNA Sequences by Unequal Crossover. *Science* 191:528-535.
- Stallings, R.L., A.F. Ford, D. Nelson, D.C. Torney, C.E. Hildebrand, and R.K. Moyzis, 1991. Evolution and Distribution of (GT)_n Repetitive Sequences in Mammalian Genomes. *Genomics* 10:807-815.
- Sundquist, W.I. and A. Klug, 1989. Telomeric DNA Dimerizes by Formation of Guanine Tetrads Between Hairpin Loops. *Nature* 342:825-829.
- Ulla, E. and C. Tschudi, 1984. Alu Sequences are Processed 7SL RNA Genes. *Nature* 312:171-172.
- Vogel, F. and R. Rathenberg, 1975. Spontaneous Mutations in Man. *Adv. Human Genet.* 5:223-318.
- Waldman, A.S. and R.M. Liskay, 1988. Dependence of Interchromosomal Recombination in Mammalian Cells on Uninterrupted Homology. *Mol. & Cell Biol.* 8:5350-5357.
- Weber, J.L. and P.E. May, 1989. Abundant Class of Human DNA Polymorphisms Which Can Be Typed Using the Polymerase Chain Reaction. *Am. J. Hum. Genet.* 44:388-396.
- Weiner, A.M., 1980. An Abundant Cytoplasmic 7S RNA is Complementary to the Dominant-Interspersed Middle Repetitive DNA Sequence Family in the Human Genome. *Cell* 22:209-218.
- Wolf, S. and P. Perry, 1974. Differential Giemsa Staining of Sister Chromatids and the Study of Sister Chromatid Exchanges Without Autoradiography. *Chromosoma* 48:341-353.

Figure 1

Unequal crossing over by one repeat in a tandem family leads to a duplication or elimination of a repeat. Similarly, crossing over by X repeats leads to duplication or elimination of X repeats.

Figure 2 Sketch of an Alu Sequence. The 160 base pair region is similar to the 130 bp region plus an insertion of about 30bp.

Figure 3 A model for gene conversion, based on a single strand break, formation of a four strand (Holliday) junction, which may or may not lead to recombination. When one strand of the heteroduplex is corrected, gene conversion has taken place.