A Markov Model for the Assembly of Heterochromatic Regions in Position Effect Variegation

B. CLARKE†, I. MCKAY§, T. GRIGLIATTI‡, V. LLOYD¶ AND A. YUAN†

† *Department of Statistics, University of British Columbia, Canada,* §*Department of Statistics, Australian National University, Australia,* ‡ *Zoology Department, University of British Columbia, and the* ¶*Institute of Molecular Biology and Biochemistry, Simon Fraser University, Canada*

(*Received on* 22 *May* 1995, *Accepted in revised form on* 14 *February* 1996)

Here we give a mathematical model for the assembly of heterochromatic regions at the heterochromatin-euchromatin interface in position effect variegation. This probabilistic model predicts the proportions of cells in which a gene is active in cells with one and two variegating chromosomes. The association of heterochromatic proteins to form remodeled chromatin following DNA replication is mainly described by accumulation independent conditional probabilities. These probabilities are conditional on the boundary of the sites to which the proteins can bind; they give the relative attractiveness of the sites to a protein complex chosen at random from a pool of available complexes. The number of complexes available is assumed to be limited and rates of reaction are implicitly modeled by the conditional probabilities. In general, these conditional probabilities are not known, however, they can be experimentally determined.

By comparing double variegation situations to single variegation, this model shows that there may be an effect on the expression of reporter genes located near the interfaces due to different sites competing for heterochromatic proteins. In addition, this model suggests that in some cases the attractiveness of sites may change in the presence of other chemical species. Consequently, the model distinguishes between two sorts of data obtained from competition experiments using position effect variegation. The two sorts of data differ as to whether there is a change in the attractiveness of sites in addition to an effect from different sites competing for the same constituents of heterochromatin. Subject to the fact that some of its parameters are not known precisely, this model replicates data from several experiments and can give predictions in other cases.

1. Introduction

While strands of DNA are often modeled only as a sequence of nucleotides, in fact, in eukaryotes, DNA molecules are packaged in proteins. The structure comprised of DNA and these chromosomal proteins is called chromatin. Chromatin is not a uniform entity; indeed, there are two broad categories of chromatin found in most eukaryotic organisms. Originally, these two types were distinguished by their degree of staining which reflects differences in packaging and condensation. The densely staining region called heterochromatin is associated with most centromeric regions and with tightly packed chromosomes such as the *Y* chromosome and the inactivated *X* chromosome in *XX* female mammals. These heterochromatic regions are thought to remain highly condensed through the life of a cell (Heitz, 1934). By contrast, euchromatic regions are less densely stained and are less condensed. These differences suggest that heterochromatin contains sequences of DNA which are seldom transcribed whereas euchromatin contains sequences that are regularly transcribed.

§ I. McKay died during the final preparation of this manuscript. Correspondence should be addressed to B. Clarke. E-mail: bclarke@stat.ubc.ca

0022–5193/96/140137 + 19 $18.00/0

Heterochromatin generally comprises 20–50% of the chromatin of the genome of a eukaryote. However, it contains few active genes compared with the euchromatic portion of the genome. For example, about 25% of the genome of the common genetic model, the fruit fly *Drosophila melanogaster* is heterochromatic and up to 50% of the DNA sequences in this heterochromatin is comprised of satellite sequences and sequences that resemble defunct transposable elements. There is very little transcriptional activity in this portion of the genome, on a per unit basis, compared with the euchromatic portion. Thus, one of the key differences between heterochromatin and euchromatin is the way they regulate gene expression.

When a normally euchromatic segment of DNA is packaged as heterochromatin its transcription usually ceases. One of the best known examples of the relationship between packaging DNA as heterochromatin and gene silencing is *X*-inactivation in mammals. This observation is corroborated by a variety of experimental data. For instance, if a normally active gene is inserted into heterochromatin by transformation it is often inactive.

Over 20 years of research by numerous laboratories has demonstrated that chromatin is a complicated entity. It is known that DNA is wound around nucleosomes which are octomers of four histones (two molecules of each of *H2A*, *H2B*, *H3*, and *H4*) and that a fifth histone, *H1*, appears to associate with the DNA linking nucleosomes. The nucleosome structure is fairly uniform throughout the genome suggesting it has little to do with the distinction between heterochromatin and euchromatin *per se*. However, histones represent only about 50% of the chromosomal proteins; the other 50% of the proteins associated with chromatin remain more enigmatic (Wolffe, 1994; Elgin, 1995). Other common chromatin proteins such as HMG1 and HMG2 (high mobility group proteins) preferentially associate with DNA regions that are thought to be kinked or distressed, while proteins such as MeCP1 are associated with methylated cytosine bases in CpG islands and these modifications appear to play a role in gene silencing (Boyes & Bird, 1991, 1992; Giese *et al*., 1992; Paull *et al*., 1993).

However, this is not the whole story either. We know that cis-acting transcription enhancing elements (DNA sequences) or locus control regions are often located thousands of base pairs away from the gene whose action they influence. Chromatin structure, mediated by protein-protein interactions, may be important in bringing these regions into close proximity with the gene promoter region. Finally, beyond packaging the DNA into the nucleosome and arrays of nucleosomes into a chromatin fibre, the chromatin is not simply heaped within the nucleus, rather, chromatin and chromosomes appear to occupy specific domains within the nucleus. This intranuclear compartmentalisation may occur as a consequence of specific associations between chromatin domains and the nuclear matrix.

Alterations in chromatin structure are the essential first step in gene expression. Chromatin domains must be relaxed, or "opened up", to allow transcription factors access to the DNA located therein so that the appropriate gene can be transcribed. Moreover, we know that DNA is replicated as chromatin. The proteins may be displaced locally but they are not stripped off long sections of DNA. Chromatin formation on the newly replicated molecules occurs very rapidly as the replication fork moves down the DNA molecule. Thus, the DNA-protein and protein-protein interactions of chromatin are clearly important in both cellular transmission of the hereditary material and in gene expression. Yet, we know little about the details of the assembly of chromatin following DNA replication or the regulation of chromatin domains in various tissue types (Wolffe, 1994).

A number of laboratories have used the phenomenon of position effect variegation (PEV) in a variety of organisms to study chromatin structure and how this structure regulates gene expression. PEV occurs when a normally euchromatic segment of DNA is placed near a broken segment of heterochromatin by an inversion, translocation, or transposition event. The normal gene (or genes) is expressed in some cells of the tissue in which it should be expressed and is repressed in others; hence the name position effect variegation. Gene silencing in PEV is correlated with the packaging of the normally euchromatic segment of DNA as heterochromatin. Both occur in about the same proportion of cells. In those cases where gene function and chromosome morphology can be observed in a single tissue, there is a precise correlation between repression of gene function and condensation (Henikoff, 1981).

For instance, in the inversion strain, $In(1)white^{mottled4}$, hereafter referred to as w^{m4}, the $white^{+}$ gene (w^{+}), which is necessary for the deposition of pigment in the pigment cells of the eye and other tissues, is placed within about 25 kb of a broken segment of *X* chromosome heterochromatin (Tartof *et al*., 1984). This gene is now expressed in about 10–15% of the eye pigment cells and repressed in others. The expression appears as a few (one to three or so) patches of red tissue on a background of colorless cells (white). This phenotype suggests that the transcriptional fate of a variegating gene is generally determined several cell divisions prior to its transcription. Once made, this decision appears to be propagated with good fidelity. That is, cells in which the gene is repressed generate daughter cells in which the gene remains repressed and cells in which the gene is transcriptionally competent generate more cells in which the gene remains transcriptionally competent and may eventually be expressed. We note that not all euchromatic regions are transcribed and, strictly speaking, not all transcription is limited to euchromatin. However, heterochromatic transcription is sparse in comparison.

Currently, the gene inactivation associated with PEV is best explained by the spread of heterochromatin across the newly formed heterochromatin/euchromatin boundary. When the heterochromatin extends far enough the reporter gene is suppressed, otherwise the reporter gene is active. Variegation, or cell-to-cell variability, is observed because the heterochromatin does not always extend far enough to encompass the reporter gene in every cell. Although the precise molecular basis of PEV is not currently known, it is known that PEV does not result from somatic mutation, gene loss, or the insertion of inactivating sequences.

In *Drosophila*, over 300 mutations that regulate variegating gene expression by altering chromosome and chromatin structure have been identified. The Su(var) mutations are dominant mutations that suppress the gene inactivation associated with PEV and alter chromosome and chromatin structure in the variegating segment of DNA, (Sinclair *et al*., 1983; Wustmann *et al*., 1988; Hayashi *et al*., 1991; Grigliatti, 1991). These loci may encode non-histone chromosomal proteins or chromosome/chromatin assembly or modifying factors, particularly those that encode components of heterochromatin. Furthermore, E(var) mutations which enhance the gene inactivation associated with PEV may encode a similar set of proteins, particularly those associated with euchromatin or transcription factors.

It is possible to conduct "bivariate" experiments in which two different variegating reporter genes, on different chromosomes, undergo PEV in the same strain. This situation can be compared with the two "univariate" experiments in which only one of the reporter genes variegates. The fraction of cells in which a reporter gene is active in the bivariate experiment can be the same as the fraction of cells in which it is active in the univariate experiment. More typically, the two fractions are different. Since there are two variegating genes, there are two pairs of fractions to compare. When no difference is observed in either pair of fractions, the two univariate variegations are phenotypically equivalent to the single bivariate variegation. This suggests that the two heterochromatic regions in the bivariate setting are comprised of non-overlapping species of proteins and may variegate independently. If a difference is observed, a possible inference is that the dependence between the two heterochromatic regions is caused by their having at least one species of protein in common. In either case, there may be other factors that result in dependence.

It has been observed that genes are more likely to be suppressed in the bivariate experiment than in the univariate experiments (Lloyd, 1995; Lloyd *et al*., 1996). This may be due, in part, to competition amongst the sites at which the chromatin molecules can bind. Competition amongst binding sites arises because the proteins in chromatin are available in limited, indeed perhaps fixed, quantities for which the sites compete. Consequently, comparing bivariate and univariate experiments can help identify regions of chromatin that contain the same species of chromosomal proteins. In addition, we find here that it may be necessary to assume that the affinities of protein molecules for sites is dependent on the range of sites to which they can bind and the boundary conditions at those sites. The term affinity is used here to mean the attractiveness of a site to a heterochromatic protein complex in the sense of likelihood of binding under certain physical conditions. We note that this is different from the term as used by chemists and biochemists to describe binding constants between two interacting molecules.

In principle, one can assume a degree of overlap in chemical species between the two univariate cases and then test those assumptions by making predictions from a bivariate experiment for its univariate experiments. The model presented here permits such testable predictions. Extensions of this model can also make predictions on how various Su(var) mutations might act alone or in combination with a second non-allelic Su(var) mutant gene in either singly or doubly variegating strains. This is essential for predicting the types of gene-protein interactions that will be seen from further genetic studies.

The model here focuses on the assembly of chromatin at an artificially created interface between heterochromatin and euchromatin. This can be caused by segmental inversion or translocation. No mathematical models describing this situation currently exist. The present model is a hierarchical Markov chain which can be applied to bivariate-univariate comparisons. It generates probabilities which describe the fraction of cells in which either or both reporter genes are suppressed. We have implemented the model computationally by a simulation procedure that mimics the real biology. This generates predictions which can be matched to experimentally obtained results. For a specific bivariate-univariate comparison we have matched existing data. Moreover, we describe later how this model can be used to make further predictions to be tested.

A qualitative model for position effect variegation based on chemical kinetics has been given by Locke *et al*. (1988). This model explains certain phenomena such as how alterations in the number of genes encoding a chromatin protein might drive reactions by altering the concentration of one or more components. However, it is difficult to test this model because it does not make precise predictions readily. The corresponding drawback to the model presented here is that it requires many input values which have not been accurately measured yet. This will typically be the case in most mathematical modeling problems in biology because the number of quantities that can be measured is dwarfed by the number of factors known to influence those quantities. It is only by assuming such knowledge that our model does give testable predictions. In particular, the model presented here requires knowledge of the relative affinity of chemical species for each other in the nucleus of a cell. Here, we have used ranges of values obtained by analogy with similar experiments where available, and have otherwise matched a specific data set so as to find plausible values. Although we have not formally done so, our model can use data to estimate these chemical affinities. Thus, the values we have used can also be tested. Constructing mathematical models can help guide the search to identify other species of chromosomal proteins and can help organise data that has been obtained. Resulting refinements may yield better predictions.

The structure of this paper is as follows. In Section 2, we describe the model and its physical basis. In Section 3, we implement the model computationally and verify that under reasonable assumptions our model matches data given in Lloyd *et al*. (1996). Finally, in Section 4 we discuss the implications of the data matching in Section 3 and the modeling strategy.

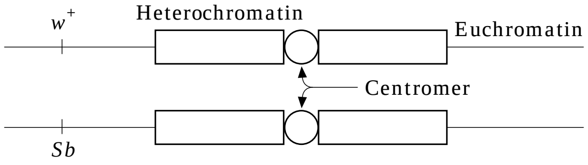

FIG. 1. Two chromosomes prior to segmental inversion. The centromeres are denoted by circles. Boxes encasing regions next to the centromeres indicate heterochromatin. Parts of chromosomes indicated by straight lines rather than encased in boxes are euchromatic. The locations of the w^+ and Sb reporter genes relative to centromeric heterochromatin are also indicated.

2. The Model

2.1. PHYSICAL BASIS OF THE MODEL

Basically, our model is an abstraction of the following example. In *Drosophila*, two reporter genes that have been used are the w^+ gene and the translocation Stubble variegator gene, $T(2;3)Sb^v$, which we abbreviate to Sb. When active, w^+ makes a fly's eyes red rather than white and Sb makes a fly's bristles short and thick rather than long and thin. These genes are on the X and third chromosomes respectively. As indicated in Fig. 1 each is located in a euchromatic region far from a heterochromatic region. Thus as shown in Fig. 2, a segment containing either of them can be inverted so as to put the reporter gene next to centromeric heterochromatin. Since the repression of w^+, for instance, is accomplished by the probabilistic propagation of heterochromatin beyond the breakpoint, w^+ is not always repressed. The eyes, therefore, have a mottled appearance composed of patches of white cells and red cells. The mottled large spot mosaic appearance suggests that the transcriptional activity of the w^+ gene is determined during embryogenesis and the decision made at that early stage is propagated with reasonable accuracy through later mitotic divisions.

If both segments, one for w^+ and one for Sb are

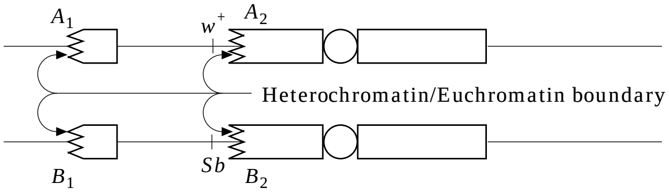

FIG. 2. Two chromosomes, A and B, which exhibit PEV. On the A chromosome there are two sites, A_1 and A_2. On the B chromosome there are two sites, B_1 and B_2. The location of each reporter gene is now between two regions of heterochromatin, as indicated by the boxes. They are closer to A_2 and B_2 than to A_1 and B_1. The ragged edges of the boxes area indicate the breakpoint for inversion. Note that the small heterochromatic regions have been broken off of sites A_2 and B_2 so as to form A_1 and B_1.

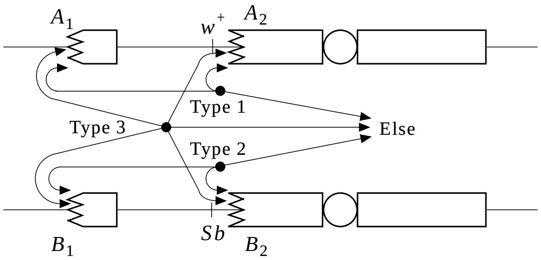

FIG. 3. Binding properties of protein complexes. The binding properties of type 1, 2, and 3 complexes are indicated by arrows. Type 1 complexes can bind to A_1, A_2 and *else*. Type 2 complexes can bind to B_1, B_2 and *else*. Type 3 complexes can bind to A_1, A_2, B_1, B_2 and *else*. *Else* indicates all heterochromatin not at the distinguished A and B sites. When enough complexes have bound to A_2, or B_2, the corresponding reporter gene is suppressed.

inverted, then there are four distinguished sites of altered heterochromatin formation. The four distinct sites correspond to the two heterochromatic regions proximal to each reporter gene. As in Fig. 2, two of these are A_2 for w^+ and B_2 for *Sb*. In addition, a segment of heterochromatin has been broken off and shifted to a distal position on the chromosome, A_1 and B_1. The rest of the chromosome set also contains possible sites for the binding of chromatin proteins associated with sites A_1, A_2, B_1 and B_2 and we call these other chromatin protein binding sites *else*.

Although there are many species of chromosomal proteins, we group them into three classes for the purpose of analysis. In fact, although we have called them protein molecules, they are likely protein complexes, composed possibly of a large collection of distinct protein species; we grouped them together here only because they have similar binding properties. The three types of protein complexes and their binding sites are as follows. As shown in Fig. 3, type 1 chromosomal proteins bind to A_1, A_2 and *else*. Type 2 chromosomal proteins bind to B_1, B_2 and *else*. The sites A_1 and A_2 are associated to w^+; the sites B_1 and B_2 are associated to *Sb*. These two types of complexes correspond to two non-interacting variegators. Competition between the *A* sites and the *B* sites is provided by chromosomal proteins of type 3 which are permitted to bind to all five sites, A_1, A_2, B_1, B_2 and *else*.

These protein complexes are available in limited amounts. They are not significantly replenished during the life of a cell, although the amount is doubled for cell division. The protein complexes are assumed to float freely in the nucleus of the cell until they bind to an available site. In fact, the nucleus is not a well-stirred solution as our model assumes. The nucleus probably consists of a collection of domains or fields and it is known that chromosomes are organised within nuclei in a non-random fashion (Hochstrasse *et al*., 1986). However, locally it may be close enough to a well-stirred solution that this assumption is valid and small inhomogeneities in distribution may be incorporated into the accumulation independent probabilities to be introduced shortly. It is possible that some complexes may be released from areas near to the breakpoints as a consequence of the disturbance in topology, but this amount is assumed to be negligible. It is also possible that the *A* and *B* sites draw protein from the rest of the chromatin. We model this as negligible also.

We model the assembly of chromatin in a cell at an early stage of the development of the embryo; in fact, the nuclear division in which chromatin is reconstructed from a uniformly staining entity into one in which euchromatic and heterochromatic regions are observed. This provides an interpretation for the probabilities generated by the model. For instance, if we model an inversion of the w^+ gene in the first cells from which all later eye cells form then all of the later eye cells will be either red or white. However, because the spread of heterochromatin is probabilistic, the proportion of flies with white eyes will equal the probability that heterochromatin covers the w^+ gene. Clearly, the fidelity with which this initial decision is propagated could influence these probabilities.

The model produces theoretical probabilities for each of the four events of interest. There are two genes, each may be active or repressed. Thus there is a 2-by-2 table of probabilities. The first two probabilities are $P(w^+$ repressed, *Sb* repressed) and $P(w^+$ repressed, *Sb* active); they sum to give the marginal probability that w^+ is repressed. Repression of a gene is equivalent to that gene having been covered by heterochromatin. The other two probabilities are $P(w^+$ active, *Sb* repressed) and $P(w^+$active, *Sb* active); they sum to give the marginal probability that w^+ is active. Similarly, we can obtain the marginal probabilities $P($*Sb* repressed) and $P($*Sb* active). The experimental data we have only gives estimates of these four marginal probabilities. Estimates of all four probabilities in the 2-by-2 table could not have been obtained for this pair of reporter genes.

The model itself is a discrete time, discrete space Markov chain. During each time step two processes occur. First, one molecule or complex of one type is selected at random from the available pool. Second, the chosen complex binds to one of its appropriate sites. This is repeated until all the available complexes are used. Random selection means that complexes bind in accordance with their prevalence—the more complexes there are of a type the more likely it is that a complex of that type is to be chosen. The influence of concentration on assembly can be modeled by standard mass action models of chemical reactions (Locke *et al*., 1988). Once a complex has been chosen in a time step, where it binds will be determined in part by boundary conditions at the sites where it can bind. That is, the boundary conditions determine the probability of binding at a site. Consequently, we assume that once bound at a site, a complex will stay bound there.

The key step in the model is the generation of probabilities at one time step in the future dependent on probabilities in the present time step. This is accomplished primarily by what we call accumulation independent conditional probabilities, AICP's. Basically, an AICP is the conditional probability that a protein complex binds to one of its sites given the boundary conditions at all of the sites to which it can bind. These represent the likelihood of a complex binding to a site which we call its affinity. The boundary conditions are defined by types of the complexes most recently bound and previously bound to each of the four sites. The AICP's are accumulation independent in that the numerical values assigned to the individual probabilities are idealised descriptions of the behaviour of the protein complexes incorporating all information except the boundary conditions. In particular, all of the AICP's are assumed to be conditioned on the same information regarding inhomogeneity of the distribution of the molecules, and influences from other molecules which are already bound to the DNA or are present in the cytoplasm. The model is Markov because the dependence in the AICP's only extends one time step back.

In reality, this process is not Markov because dependency extends backwards in time to the very beginning. Markovity only captures what is perhaps the strongest dependency. More experimentation is required before more detailed modeling can be justified. Indeed, a plausible source for a form of dependency that Markovity precludes arises from heterochromatin nucleation sites. We pursue this discussion in Section 4.

Once all of the complexes have been allocated, the activity of the two genes is determined by how many complexes have bound to A_2 and B_2. Thus, we use cut-off values M_1 and M_2. If at least M_1 complexes have been bound at A_2, then the w^+ reporter gene is inactive. Otherwise, the gene is fully active. Similarly, if at least M_2 complexes have bound at B_2, the *Sb* reporter gene is inactive. Consistent with what is observed biologically, this model does not permit intermediate levels of activity.

Since the allocation of complexes to sites is a probabilistic process we can simulate arbitrarily many independent iterations of it. This generates estimates of the four probabilities of interest as the fraction of times over all iterations that the genes are repressed or active. Provided the number of repetitions is sufficiently large these probabilities can be known to any desired degree of accuracy. However, experimental results can only achieve a finite degree of accuracy so the mathematical precision cannot be fully tested. The probabilities the simulation generates depend on the number of complexes of each type, the cut-off values and the AICP's. Although none of these quantities is known with certainty, the number of unknowns can be reduced substantially by symmetry considerations and ranges of plausible values can be established by comparison with other experiments. Moreover, all of these unknowns are natural: they describe real phenomena and they can be measured. This approach identifies basic quantities which appear to govern position effect variation. Thus, it may motivate experimental testing of the exact values we have used.

2.2. FORMAL PRESENTATION OF THE MODEL

To formalise this model we recall that Markov chains are defined by their initial probability distribution and their transition matrices. We begin by defining a state vector which will describe the system for each time t. Then, since the initial states are known, it is enough to specify the time evolution of the probabilities. Let P_i denote the number of complexes of type i present in the nucleus at time $t = 0$. So, initially, there are $N = P_1 + P_2 + P_3$ complexes in total and the Markov chain runs for N time steps. Implicitly this assumes that all reactions proceed at about the same rate; small differences can be accommodated by the AICP's.

The state of the system after t time steps is given by a vector of 15 non-negative integers, written as a time dependent vector of two subvectors of length 11 and four. The first subvector gives the number of complexes at each site of each type that have been bound. The second gives the boundary conditions; it identifies the type of complex most recently bound at the four distinguished sites. *Else* does not have boundary conditions and is presumed to be large relative to the other four sites. We write this vector as

$$(X(t); C(t)) = (X_1(t), \ldots, X_1 1(t);$$
$$C_1(t), C_2(t), C_3(t), C_4(t)),$$

where

$X_1\ (t) =$ number of type 1 complexes attached to A_1 at time t

$X_2(t) =$ ....3..................................A_1

$X_3(t) =$ ....1..................................A_2

$X_4(t) =$ ....3..................................A_2

$X_5(t) =$ ....2..................................B_1

$X_6(t) =$ ....3..................................B_1

$X_7(t) =$ ....2..................................B_2

$X_8(t) =$ ....3.................................. B_2

$X_9(t) =$ ....1..................................*else*

$X_{10}(t) =$ ....2..................................*else*

$X_{11}(t) =$ ....3..................................*else*

and

$$C_1(t) = \begin{cases} 0, \text{ if at time } t \text{ the last complex attached to } A_1 \text{ is type 1;} \\ 1, \text{ if last complex attached at } A_1 \text{ is type 3.} \end{cases}$$

$$C_2(t) = \begin{cases} 0, \text{ if at time } t \text{ the last complex attached to } A_2 \text{ is type 1;} \\ 1, \text{ if last complex attached at } A_2 \text{ is type 3.} \end{cases}$$

$$C_3(t) = \begin{cases} 0, \text{ if at time } t \text{ the last complex attached to } B_1 \text{ is type 2;} \\ 1, \text{ if last complex attached at } B_1 \text{ is type 3.} \end{cases}$$

$$C_4(t) = \begin{cases} 0, \text{ if at time } t \text{ the last complex attached to } B_2 \text{ is type 2;} \\ 1, \text{ if last complex attached at } B_2 \text{ is type 3.} \end{cases}$$

The X_i's count the number of protein complexes of each type that are bound to each site; at time $t = 0$ all X_i's are zero since no complexes have been bound. Since one protein complex is bound at random during each time step, the time t is the total number of protein complexes that have been bound by that time to the possible sites A_1, A_2, B_1, B_2 and *else*. Thus, $t = \Sigma_{i=1}^{11} X_i(t)$. The vector $\mathbf{C}(t)$ is the boundary condition at time t since it identifies the type of the most recently bound complex. The AICP's for each time step therefore depend on $\mathbf{C}(t)$. At $t = 0$, there will be 16 possible initial values for $\mathbf{C}(t)$ to assume; the program we describe later permits these to vary according to a uniform distribution over the 16 possibilities.

The gene w^+ is inactivated at time t if and only if it has been covered by heterochromatin at time t. That is, if and only if at least M_1 units of protein complex types 1 and 3 have bound to A_2. Equivalently, if and only if

$$X_3(t) + X_4(t) \geqslant M_1.$$

Likewise, *Sb* is inactivated at time t if and only if M_2 units of protein complex types 2 and 3 have bound to B_2. That is

$$X_7(t) + X_8(t) \geqslant M_2.$$

The constants M_1 and M_2 must be chosen in advance, based on the known approximate linear length between genes and heterochromatin. The constants M_1 and M_2 represent the minimal number of complexes required to be bound to sites A_2 and B_2 in order to preclude transcription. Reporter genes are fully active until M_1 or M_2 complexes have been bound respectively, and after that the gene is completely suppressed.

Explicit expressions for the probability of selecting a type of complex and letting it bind to one of its sites are given in the Appendix. For the present, we comment that there are 24 AICP's which govern the binding of the complexes. Each AICP corresponds to a collection of free parameters of the form

$$P(\text{a type } i \text{ goes to a particular site} | (c_1, \ldots, c_4))$$

whose values must be assigned, see Table 1 for a listing. We refer to the collection of such values for a fixed set of boundary conditions as an AICP and we refer to each member of such a collection as an AICP. This terminology is standard in probability theory as the context makes it clear which meaning is intended.

The collection of AICP's can be enumerated. This is shown in Table 1. There are 4 × 3 = 12 such numbers for type 1 and that many again for type 2; 24 in all. There are 16 × 5 = 80 such numbers for type 3 and so 104 in total. However, not all of these are free parameters; the number of free parameters can be reduced substantially.

To reduce the number of free parameters we impose three physically meaningful criteria. First, for each fixed boundary condition the values must sum to 1 so that each is a valid probability. Second, we regard the probability of a protein complex going to *else* as the probability that a complex goes to another of the same type situated away from the distinguished sites. This is permissible because we assume that complexes of a given type are most attracted to complexes of the same type or of an appropriate type based on chemical reactions, and *else* is so big that it contains replicating units and chromatin domains of virtually all types. We ignore complexes in *else* that are not of the same type as the one being added on the grounds that because *else* is so large they will contribute negligibly. Operationally, this means that for any set of boundary conditions that contains a type i, the probability of a type i going to *else* as given in Table 1 will be the same as the maximum of the other entries on that row. This constrains the probability of a type going to *else* in all rows except for rows 4, 8, and 9, the rows for which the boundary values do not contain a complex of the appropriate type for the protein–protein interaction. There are three such rows because there is one row for each type. Third, we impose symmetry between comparable sites. That is, if two sites have the same boundary condition then they have the same probability of binding an incoming complex.

Imposing these three conditions on the entries of Table 1 reduces the apparent degrees of freedom as follows. Row 1 has no free parameters; *else* is regarded as a type 1, so by symmetry all sites are equiprobable. To satisfy the definition of a probability the three entries also sum to one. Now, there are three constraints so all three unknowns are determined. In row 2, there is one free parameter. The entries sum to one and by the assumption on *else* it is the maximum of the first two entries. Three unknowns with two constraints leaves one free parameter. Row 3 is just a permutation of row 2, by symmetry, and so introduces no further free parameters. In row 4, by symmetry, two entries are the same and the entries sum to one. This leaves one free parameter. Rows 5–8 have 0,1,0,1 free parameters respectively, by the same reasoning. Thus, the first eight rows only have four free parameters.

Rows 9–24 are more complicated because they have five entries each. In row 9, the first two entries are equal and the second two entries are equal. The entries also sum to one. Three constraints on five parameters leaves two free parameters. In row 10, the second two entries are equal, and the maximality of *else* means that the fifth entry equals the first entry. Since the entries sum to one there are only two free parameters. Row 11 is a permutation of row 10 and so introduces no further free parameters. Row 12 has its first two entries equal and its second two entries equal. The first two entries equal the fifth by maximality and the sum of the entries is one. This

TABLE 1
Theoretical values for the AICP's

Boundary	Site A_1	Site A_2	Site B_1	Site B_2	*Else*
(1,1,0,0)	x	x	0.000000	0.000000	x
(3,1,0,0)	$z < x$	x	0.000000	0.000000	x
(1,3,0,0)	x	$z < x$	0.000000	0.000000	x
(3,3,0,0)	z	z	0.000000	0.000000	$x > z$
(0,0,2,2)	0.000000	0.000000	y	y	y
(0,0,3,2)	0.000000	0.000000	$z < y$	y	y
(0,0,2,3)	0.000000	0.000000	y	$z < y$	y
(0,0,3,3)	0.000000	0.000000	y	y	$z > y$
(1,1,2,2)	x	x	y	y	$z > x > y$
(3,1,2,2)	z	x	y	y	$z > x > y$
(1,3,2,2)	x	z	y	y	$z > x > y$
(3,3,2,2)	z	z	y	y	$z > y$
(1,1,3,2)	x	x	z	y	$z > x > y$
(3,1,3,2)	z	x	z	y	$z > x > y$
(1,3,2,2)	x	z	z	y	$z > x > y$
(3,3,3,2)	z	z	z	y	z
(1,1,2,3)	x	x	y	z	$z > x > y$
(3,1,2,3)	z	x	y	z	$z > x > y$
(1,3,2,3)	x	z	y	z	$z > x > y$
(3,3,2,3)	z	z	y	z	$z > y$
(1,1,3,3)	x	x	z	z	$z > x$
(3,1,3,3)	z	x	z	z	$z > x$
(1,3,3,3)	x	z	z	z	$z > x$
(3,3,3,3)	z	z	z	z	z

This table shows the theoretical AICP's. There are four for each of type 1 and type 2; there are 16 for type 3. The left most column shows the 24 relevent boundary conditions. The columns correspond to the sites to which complexes can bind. In each row a variable is used to indicate the theoretical value and its size relative to the other entries is indicated; x is used for type 1, y for type 2 and z for type 3. These placeholders are not in general the same from row to row. Further symmetries in the Table are discussed in the text. A string of zeroes indicates that the entry does not appear.

leaves one free parameter. These rows therefore give 2,2,0,1 free parameters respectively.

Row 13 has two free parameters because the first two entries are equal, as are the fifth and third; and the entries sum to one. In row 14, the first, third and fifth entries are the same. Since the entries sum to one, there are two free parameters. Row 15 is a permutation of row 14, so no further free parameters. Row 16 has the first three entries equal and they equal the fifth entry. Together they sum to one so there is one free parameter. These rows give 2,2,0,1 further free parameters. The boundary conditions on rows 17–20 are just a permutation of the boundary conditions on rows 12–15. So, rows 17, 18, 19, and 20 are determined by rows 13, 14, 15, and 16, respectively. So, none of these four rows add any extra parameters. Thus, rows 13–20 add 2,2,0,1,0,0,0,0 free parameters, respectively.

Row 21 has its first two entries equal, and its last three entries equal. They sum to one so there is one free parameter. Row 22 has its first entry equal to each of the last three. Summing them to one means there is one free parameter remaining. Row 23 is a permutation of row 22. Row 24, like rows 1 and 5, has no free parameters. These rows give 1,1,0,0 free parameters respectively. Thus, the total number of free parameters over all 24 rows is the sum of the numbers of the free parameters we have identified, namely, 16. Table 2 shows a specific choice for the entries of Table 1. This choice satisfies all of the above constraints.

For ease of interpretation we actually multiply the probabilities of going to *else* by a constant ξ which we call the size of *else*. The larger the ξ, the more attractive is *else*. In particular, one can interpret its role in the formulae as follows. The AICP's in Table 1 show the probabilities assuming two "A" sites, two "B" sites and one *else* site. If in fact there were, for example, ten *else* sites then all of the AICP's for going to *else* should scale up in the same way, regardless of type. This does not imply all types are equiprevalant in *else* because the AICP's may in fact be different. Moreover, this is consistent with our reduction of *else* to a site with boundary condition the same as the next complex to be bound because in the limit of large *else* it is only the most attractive sites that need to be considered; the other sites contribute ever less attractiveness proportionally as *else* increases.

One can interpret the size of *else* in terms of the AICP's. Recall that type 1 complexes can go to A_1, A_2 or to *else*. For a set of boundary conditions, write the corresponding AICP's as p_1, p_2 and p_3. Now, if ξ is the size of *else*, the conditional probabilities become p_1/S, p_2/S and $\xi p_3/S$ where $S = p_1 + p_2 + \xi p_3$ is a normalising constant. Thus, ξ is an adjustment factor reflecting how many more sites there are in *else* than in the four distinguished sites. Roughly, if there are ten times as many *else* sites as other sites, the attraction of *else* goes up by a factor of ten when an incoming molecule has to bind somewhere. We comment that *else* not precisely defined physically. It may not reflect the whole genome apart from the four sites, it may only reflect the attraction of parts of the genome that are close enough to the four distinguished sites that a specific protein complex has a chance of going there.

This means we have a total of 16 free parameters for the AICP's, one more for the size of *else* and five more parameters P_1, P_2, P_3, M_1 and M_2 to count molecules or protein complexes. Thus, there is a total of 22 parameters that must be specified in order to obtain the probabilities of interest from simulations of the model. In Section 3.3 we shall return to the problem of specifying the parameter values.

3. Implementation of the Model

3.1. DESCRIPTION OF THE PROGRAM

The four relevant probabilities for the expression or non-expression of genes are easily determined by means of simulation, that is, by Monte Carlo techniques. We did this by a program written in the "c" programming language, compiled and run on a Sun Microsystems SPARC model 10 workstation.

Our program accepts as inputs AICP's as in Tables 1 and 2, and the additional parameters ξ, M_1 M_2, P_1, P_2, P_3. For convenience, the program sums the entries in each AICP row and then divides the entries of the row by this sum so that any set of positive numbers can be entered for normalisation to give a probability. In principle, the program is able to reproduce and save an arbitrary number of independent instances of the entire simulation process, although this is unnecessary for the purpose of estimating the probabilities in question. Thus, in practice the central routine is called from within another program that provides a more "user-friendly" interface and stores only the requested information, namely the estimated probabilities.

The various parts of the program described below, are integrated by a "main" function which reads parameters from the command line and inputs the AICP's from a file. Main also modifies the AICP's using the "size of *else*" parameter ξ and runs the simulations. The results of the simulations can be output to a file nominated by the user. The output of the program is a 2-by-2 contingency Table.

TABLE 2

AICP's for the bivariate $w^+ - Sb$ variegator

Row	Boundary	Site A_1	Site A_2	Site B_1	Site B_2	*Else*
1.	(1,1,0,0)	0.333333	0.333333	0.000000	0.000000	0.333333
2.	(3,1,0,0)	0.300000	0.350000	0.000000	0.000000	0.350000
3.	(1,3,0,0)	0.350000	0.300000	0.000000	0.000000	0.350000
4.	(3,3,0,0)	0.333333	0.333333	0.000000	0.000000	0.333333
5.	(0,0,2,2)	0.000000	0.000000	0.333333	0.333333	0.333333
6.	(0,0,3,2)	0.000000	0.000000	0.220000	0.390000	0.390000
7.	(0,0,2,3)	0.000000	0.000000	0.390000	0.220000	0.390000
8.	(0,0,3,3)	0.000000	0.000000	0.250000	0.250000	0.500000
9.	(1,1,2,2)	0.250000	0.250000	0.110000	0.110000	0.300000
10.	(3,1,2,2)	0.300000	0.200000	0.110000	0.110000	0.300000
11.	(1,3,2,2)	0.200000	0.300000	0.110000	0.110000	0.300000
12.	(3,3,2,2)	0.266666	0.266666	0.110000	0.110000	0.266666
13.	(1,1,3,2)	0.200000	0.200000	0.250000	0.110000	0.250000
14.	(3,1,3,2)	0.233333	0.200000	0.233333	0.110000	0.233333
15.	(1,3,2,2)	0.200000	0.233333	0.233333	0.110000	0.233333
16.	(3,3,3,2)	0.225000	0.225000	0.225000	0.110000	0.225000
17.	(1,1,2,3)	0.200000	0.200000	0.110000	0.250000	0.250000
18.	(3,1,2,3)	0.233333	0.200000	0.110000	0.233333	0.233333
19.	(1,3,2,3)	0.200000	0.233333	0.110000	0.233333	0.233333
20.	(3,3,2,3)	0.225000	0.225000	0.110000	0.225000	0.225000
21.	(1,1,3,3)	0.150000	0.150000	0.233333	0.233333	0.233333
22.	(3,1,3,3)	0.220000	0.120000	0.220000	0.220000	0.220000
23.	(1,3,3,3)	0.120000	0.220000	0.220000	0.220000	0.220000
24.	(3,3,3,3)	0.200000	0.200000	0.200000	0.200000	0.200000

Each row of the table shows an AICP, where the boundary is specific in the second column. A zero indicates that the conesponding boundary does not affect the AICP. These AICP's were used in matching the bivariate case. They satisfy all the constraints given in the text.

To take full advantage of the function handling features of the c language, our program is broken into a number of small and functionally independent pieces. The most basic component for the purpose of simulation is the system function "drand48" for generating pseudo-random sequences of uniform variables. It could be replaced with any other compatible function. The seed value for any given sequence is defined in a header file, and can be easily changed, though we did not find this necessary in most runs.

The drand48 system function is used to drive another function "discrete-samp" which samples from a discrete distribution with given probabilities; from this we are able to generate one time step if we are given the current state vector, the numbers of molecules or protein complexes of different types currently available, and the corresponding conditional probabilities. All of this is done in a function called "conditional-samp". (For ease of maintenance, the work of calculating conditional probabilities is performed in another function called "conditional-prob", which we describe below.) From these components, it is easy to simulate one instance of the process within an iterative loop. This is done in a function called "samp-path", which returns the "hitting times", ie. the time steps at which the genes are inactivated, if at all. It is only necessary to choose boundary conditions randomly to initiate the simulation; this is done with a function called "random-boundaries".

Computation of conditional probabilities is conceptually a separate problem from simulation, and for this reason we compute them in a separate suite of functions. Most of the work is done in the function "conditional-prob" mentioned above, and is straightforward. Nonetheless, we point out that in Section 4 we discuss an extension of the model given in Section 2. This extension includes a "reverse gravity" feature. Briefly, reverse gravity " corrects" conditional probabilities to account for what has happened before the current time step in a non-Markovian way. If this correction is desired, the computation is done in a separate module. To avoid round-off and underflow problems we compute odds factors with which to update AICP's. These odds factors, which are simply ratios of reverse gravity effects, are imported into the function conditional-samp and the updating is done there. Finally, additional factors for availability of molecule or complex types are computed and a vector of conditional probabilities is returned to the main program.

As a test that the model and program perform the way they should, we verified that for various choices of inputs (ξ, AICP's, cutoffs, P_i's) the probabilities generated varied in an appropriate manner. For example, coverage probabilities, which measure the "spread" of heterochromatin at particular interfaces, increase as availability of complexes increases and decreases as cut-offs increase or as ξ increases. The assembly of heterochromatin at one boundary and thus the coverage probabilities for the variegation reporter gene also increased or decreased as the entries in the AICP table made a site more or less attractive to incoming complexes.

3.2. DESCRIPTION OF THE EXPERIMENTS

The data we have used appears in Lloyd *et al.* (1996, table 2). Specifically, we have matched the data obtained for male flies using the w^{m4} allele of the w^+ gene and the *Sb* allele of the *Stubble* gene. These experiments gave results as follows. First, for the bivariate case of w^{m4} and *Sb* as a double variegator, the marginal coverage probabilities for the two genes were found to be 0.77 ± 0.05 and 0.44 ± 0.05, respectively. For the univariate w^{m4} variegator, coverage of 0.91 ± 0.04 was observed; for the univariate *Sb* variegator coverage of 0.51 ± 0.04 was observed. Note that for both genes the probability of coverage is higher for the univariate cases than for marginals from the bivariate case.

The present model does not explicitly include quantities which can reflect modifications of the precise conditions under which the experiment was performed. However, a major factor such as sex or mutant alleles that strongly modify PEV might be modeled by using different numbers of heterochromatin complexes, different complexes (with different affinities) and differing sizes of *else*. For example, the Y chromosome which is found in males but not females is large and mostly heterochromatic; it may serve as a sink for heterochromatin complexes (Zucherkendl, 1974). Cases where the ploidy (number) of Y chromosomes is altered relative to the X chromosome and autosomes might be accommodated in the model by simply altering the size of *else*—either increasing it for XYY or XXY individuals or decreasing it for X/O males. Moreover, the present model assumes two alleles for each gene only one of which is variegating. The other is a null mutant, a non-functional allele in the normal genomic position in the case of wild-type reporter genes or a wild type allele in the normal genomic position in the case of mutant reporter genes (such as *Sb*).

Experimental results show that in some cases, heterochromatic complexes can be drawn from other places on the same chromosome, indeed possibly from other chromosomes. Thus, the pool of complexes presupposed in the present model might have to be reinterpreted as an idealisation of the complexes which may be drawn from various sources in the cell. The identity of the complexes chemically is a question we do not address here, although, we imagine that type 1 and type 2 complexes are specific non-histone chromosomal protein complexes, whereas type 3 behaves like a general non-histone component of chromatin or like a histone complex.

3.3. COMPUTATIONAL RESULTS

First, we set $M_1 = M_2 = 100$. That is, 100 protein complexes binding to either A_2 or B_2 is enough to inactivate a gene. This is not an unreasonable number, however, it was chosen arbitrarily to permit easy comparisons. The number can be altered easily with no effect provided the AICP's are altered accordingly.

Next we regarded type 3 complexes as the most prevalent. This reflects the assumption that most complexes can bind anywhere and that localisation is through a large collection of complexes that occur in smaller quantities. So, we chose $P_3 = 4000$ to represent the number complexes in the cytoplasm which can bind to diverse sites and set $P_1 = P_2 = 1000$ for the bivariate case. For univariate cases we used the same value for P_3, but set either P_1 or P_2 to zero and set the corresponding AICP entries to zero.

We require all AICP's to be between 0.05 and 0.95 on the grounds that otherwise measurements cannot be made with accuracy. This means that there will be a range of ξ so that for ξ less than this range both A_2 and B_2 are always covered and for ξ greater than this range neither A_2 nor B_2 will ever be covered. This makes mathematical sense and has been computationally verified. In the former case *else* attracts too few complexes and in the latter case it attracts too many complexes. In either case variegation cannot occur because one outcome happens with probability so close to one as to be essentially deterministic. Recalling that *else* encompasses much more of the genome than the four distinguished sites do, we expect ξ to be large, and in practice usually ξ must be between 30 and 65 for non-trivial results. Now, six of the 22 parameters have been identified or constrained. We turn next to justifying values for the remaining free 16 parameters in the AICP's.

We start by imposing further constraints to reflect plausible physical interpretations. First, we assume that each type prefers to bind to a site which has bound another molecule of the same type or of an appropriate type most recently. Different types of complexes prefer others of their own type to possibly different degrees. Second, from the fact that w^+ has a higher probability of coverage than *Sb* in both the bivariate and univariate settings we surmise that type 3 complexes prefer type 1 complexes over type 2 complexes. Third, for the first eight rows, we also surmise that type 1's like type 3's more than type 2's like type 3's. Fourth, the fact that the univariate coverage probability of *Sb* is essentially the same as its bivariate coverage probability implies that, for the last 16 rows, when a type 3 is being bound, the degree to which type 3's prefer type 1's over type 2's (while still preferring other type 3's most) is accentuated. That is, type 1's are more attractive to type 3's in the presence of type 2's. This follows from the fact that whatever AICP's are chosen, the univariate coverage probability for a gene will be higher than its bivariate coverage probability because the possibility of competition increases with the number of sites, for the values of P_1, P_2 and P_3 chosen here.

These qualitative statements provide ranges that the AICP values must satisfy. All rows of Table 2 do satisfy the range constraints these four statements imply. For instance, in row 2, the first two rows indicate that a type 1 would rather join a site which has a type 1 than a type 3. Comparing this to row 6, it is seen that type 2's like type 3's less than type 1's like type 3's. Row 15 indicates that a type 3 would prefer to bind to another type 3 but that type 1 is only a little less attractive and type 2 is about half as attractive. We note that in comparing, say, rows 12 and 13, the probability of a type 3 going to B_2 did not change, even though the boundary condition at B_1 did.

We note that a degree of asymmetry is tolerated. We expect that the attractiveness of a type 1 for a type 3 should be close to the attractiveness of a type 3 for a type 1 although they need not be exactly the same. It is permissible to regard the binding of a type 1 to sites containing types 1 and 3 as different from the binding of a type 3 to sites containing 1 and 3. There is no reason to assume symmetry here, however, the degree of asymmetry should not be too large. In addition, we permit the attractiveness of type 3's to type 1's to depend a little on the presence of type 2's, although this cannot be a large dependence because type 2's and type 3's are not very compatible. Again, the entries of Table 2 satisfy these qualitative constraints.

We note that Table 2 is not uniquely defined by these constraints. However, the entries cannot realistically be known to better than plus or minus 0.05. In practice, we start with a set of AICP values and find a value of ξ which gives coverage probabilities close to those observed in the bivariate case. Then, we fix that value of ξ and modify the AICP values consistent with the above constraints to get a better match. If all the probabilities of events in which a gene is covered are too high or too low we change the ξ value accordingly. In this way we iterate to get AICP's and a ξ value which matches the bivariate case exactly.

By this informal method we arrived at the values in Table 2. Using these AICP's and $\xi = 38.5$ our model generates probabilities in agreement with those obtained experimentally in the bivariate case. In particular, we found that the probability of neither gene covered was 0.14, that both were covered was 0.35 and the probability of only w^+ covered was 0.42, the probability of only *Sb* covered was 0.09. Thus, we have $0.42 + 0.35 = 0.77$ and $0.09 + 0.35 = 0.44$, matching the data perfectly.

We can now turn to the univariate cases. First consider w^+ and use $\xi = 38.5$ as determined earlier. Note that, experimentally, the marginal probability of coverage in the univariate case is 0.91, an increase over 0.77 from the bivariate. This is consistent with a decrease in competition. In Table 3 we show the AICP's that result from assuming that type 2's are not present. There are no AICP's for the addition of type 2's and the AICP entries for a type 3 binding to B_1 or B_2 are zero. Remarkably, our program gives a probability of coverage of 0.90666 which matches the data exactly. This was quite surprising, but appears to hold for differing numbers of iterations.

In this one special example, the bivariate model has been used to make a prediction for the univariate case that is in fact observed experimentally. We regard this as strong evidence that the model has in fact captured some real aspects of the competition between sites in the case that affinities do not change. Or, equivalently, since a model which only uses competition succeeds at explaining the data it follows that competition may be the major factor driving PEV at the two interfaces measured in this case.

Turning next to the univariate *Sb* variegator, we began by modifying the Table 2 AICP's analogously to Table 3. That is, since type 1's do not matter, we deleted the first four rows of Table 2 and then set the first two columns in the type 3 AICP's to zero since the A_1 and A_2 sites don't exist. Using $\xi = 38.5$, this gives a coverage of 0.58, in contrast to the experimentally observe 0.51. The value 0.58 assumes that only competition has been modeled. If we modify the AICP's a little i.e., assume that affinities change in the absence of type 1's then we get 0.49, again a good match. Table 4 shows the AICP's that worked in this case. The only change is that the entries 0.11 got changed to 0.10. This reflects the physical assumption that type 3's, and consequently

Table 3
AICP's for the univariate w^+ variegator

Row	Boundary	Site A_1	Site A_2	Site B_1	Site B_2	*Else*
1.	(1,1,0,0)	0.333333	0.333333	0.000000	0.000000	0.333333
2.	(3,1,0,0)	0.300000	0.350000	0.000000	0.000000	0.350000
3.	(1,3,0,0)	0.350000	0.300000	0.000000	0.000000	0.350000
4.	(3,3,0,0)	0.333333	0.333333	0.000000	0.000000	0.333333
9.	(1,1,2,2)	0.250000	0.250000	0.000000	0.000000	0.300000
10.	(3,1,2,2)	0.300000	0.200000	0.000000	0.000000	0.300000
11.	(1,3,2,2)	0.200000	0.300000	0.000000	0.000000	0.300000
12.	(3,3,2,2)	0.266666	0.266666	0.000000	0.000000	0.266666
13.	(1,1,3,2)	0.200000	0.200000	0.000000	0.000000	0.250000
14.	(3,1,3,2)	0.233333	0.200000	0.000000	0.000000	0.233333
15.	(1,3,2,2)	0.200000	0.233333	0.000000	0.000000	0.233333
16.	(3,3,3,2)	0.225000	0.225000	0.000000	0.000000	0.225000
17.	(1,1,2,3)	0.200000	0.200000	0.000000	0.000000	0.250000
18.	(3,1,2,3)	0.233333	0.200000	0.000000	0.000000	0.233333
19.	(1,3,2,3)	0.200000	0.233333	0.000000	0.000000	0.233333
20.	(3,3,2,3)	0.225000	0.225000	0.000000	0.000000	0.225000
21.	(1,1,3,3)	0.150000	0.150000	0.000000	0.000000	0.233333
22.	(3,1,3,3)	0.220000	0.120000	0.000000	0.000000	0.220000
23.	(1,3,3,3)	0.120000	0.220000	0.000000	0.000000	0.220000
24.	(3,3,3,3)	0.200000	0.200000	0.000000	0.000000	0.200000

Modification of the entries in Table 2 for the univariate w^+ case. Rows 5–8 in Table 1 have been deleted. Entries for sites B_1 and B_2 have been set to zero for the type 3 AICP's.

else, become relatively more attractive to type 3's in the absence of type 1's. That is, there is a change in affinity that must be modeled in addition to a competition effect.

Due to overparametrisation, an alternative way to model this would be to increase the number of type 2 complexes available, or decrease M_1. However, either of these would necessitate corresponding changes is the AICP's for the bivariate and w^+ cases. While this is possible, it makes comparison amongst cases difficult. The notion of a change in affinity here may reflect a hidden factor the present model has not properly taken into account. We return to these points presently.

We note that rows 8 and 9 are similar, but not identical. Row 8 says that a type 2 binds 25% of the

Table 4
AICP's for the univariate Sb variegator

Row	Boundary	Site A_1	Site A_2	Site B_1	Site B_2	*Else*
5.	(0,0,2,2)	0.000000	0.000000	0.333333	0.333333	0.333333
6.	(0,0,3,2)	0.000000	0.000000	0.220000	0.390000	0.390000
7.	(0,0,2,3)	0.000000	0.000000	0.390000	0.220000	0.390000
8.	(0,0,3,3)	0.000000	0.000000	0.250000	0.250000	0.500000
9.	(1,1,2,2)	0.000000	0.000000	0.100000	0.100000	0.300000
10.	(3,1,2,2)	0.000000	0.000000	0.100000	0.100000	0.300000
11.	(1,3,2,2)	0.000000	0.000000	0.100000	0.100000	0.300000
12.	(3,3,2,2)	0.000000	0.000000	0.100000	0.100000	0.266666
13.	(1,1,3,2)	0.000000	0.000000	0.250000	0.100000	0.250000
14.	(3,1,3,2)	0.000000	0.000000	0.233333	0.100000	0.233333
15.	(1,3,2,2)	0.000000	0.000000	0.233333	0.100000	0.233333
16.	(3,3,3,2)	0.000000	0.000000	0.225000	0.100000	0.225000
17.	(1,1,2,3)	0.000000	0.000000	0.100000	0.250000	0.250000
18.	(3,1,2,3)	0.000000	0.000000	0.100000	0.233333	0.233333
19.	(1,3,2,3)	0.000000	0.000000	0.100000	0.233333	0.233333
20.	(3,3,2,3)	0.000000	0.000000	0.100000	0.225000	0.225000
21.	(1,1,3,3)	0.000000	0.000000	0.233333	0.233333	0.233333
22.	(3,1,3,3)	0.000000	0.000000	0.220000	0.220000	0.220000
23.	(1,3,3,3)	0.000000	0.000000	0.220000	0.220000	0.220000
24.	(3,3,3,3)	0.000000	0.000000	0.200000	0.200000	0.200000

Modification of Table 2 for the univariate *Sb* case. Rows 1–4 have been deleted, entries for sites A_1 and A_2 for type 3 complexes have been set to zero, and occurences of 0.11 have been changed to 0.1; this indicates that type 2 are less attractive to type 3's in the absence of type 1's.

time to a site with a type 3. Row 9 says that a type 3 binds 20% of the time to a site with a type 2. The two percentages pertain to the binding of different complexes and differ but not by a lot. Rows 9–12 pertain to the binding of type 3 complexes only and do not have the same probability of a type 3 going to *else*. However, this only means that the effective probability of a molecule going to *else* under any of those four boundary conditions is a weighted average of the four values. Forcing these to be the same would reduce the number of parameters somewhat. We have not done this since the discrepancies are within ± 0.05, the degree of accuracy to which such quantities might be measured. Exact matching is possible. However, because of the overparametrisation, limited data, and experimental error exact matching must be regarded as artificial.

Permitting the relative attractiveness to change in the presence of other species seems to be necessary to match the data because the coverage of *Sb* does not change enough from univariate to bivariate. The reasoning is as follows. Suppose we have found AICP's which match in a bivariate case. If we set the appropriate entries to zero as in Table 3 then the coverage probability in the univariate case must increase because there are relatively more molecules per site. However, for *Sb* the observed increase is a bit less than the model predicts. Thus, there must be a physical basis to account for it. We propose a change in chemical affinities as used in Table 4. Actual chemical affinities do not change, but the likelihood of binding might since binding is influenced by chemical affinities under physical conditions and concentrations. The end result would be a change in affinities as we have used the term here. Indeed, the change is not great and so is physically plausible. However, it has a significant effect on the coverage of the *Sb* gene.

In view of the way univariate predictions are obtained from bivariate matching, it is worth noting that it may be possible to refine the present model by reducing the number of free parameters it uses. This might be done by requiring, for instance, that the probabilities of type 3 complexes going to *else* be constant for each set of *B*-site boundary conditions when the *A*-site AICP entries are set to zero. A similar constraint can be imposed for each set of *A*-site boundary conditions. We have not done this here because we found that, in practice, the range of probabilities appearing in the rightmost column of Table 1 is relatively small, in particular within experimental error. As it stands, these slight discrepancies only mean that the effective probability of type 3 going to *else* is a weighted sum of the appropriate entries. For instance, we see from Table 2 that in rows 9–12, 13–16, 17–20, and 21–24—the blocks on which the *B*-site boundary conditions are constant—the probability of *else* has a slightly different meaning from row to row as the boundary conditions change. In particular, if the first two entries of rows 9–12 are zero, the effective probability of a type 3 going to *else* is a bit less than 0.6.

Thus, for the $w^+ - Sb$ case, the bivariate model has generated predictions for the two univariate cases. The model successfully predicted univariate behaviour when affinities do not change. For the *Sb* case, the bivariate model led us to a physically plausible, and slight, modification of the bivariate AICP's to match the data. Arguably this modification is so small as to be artificial when contrasted with the standard errors on estimates of the true physical quantities. Nonetheless, we propose that models must distinguish those cases in which the coverage probabilities change solely due to competition from those cases where coverage probabilities change due to a change in the relative affinities of types in addition to competition. When the affinities do not change our model as it stands may give valid predictions. When the likelihoods of binding, or affinities, do change we expect that our model will overestimate the coverage probability, but be easy to modify to get a match.

4. Discussion

4.1 SUMMARY AND IMPLICATIONS

Here we have proposed a hierarchical Markov chain to formalise a competition model for the assembly of chromatin in PEV. Depending upon the input parameters, this model generates theoretical values for the coverage probabilities of reporter genes. The model has numerous parameters, most of which are not known. However, we reduced the number of free parameters substantially by physically plausible assumptions and constrained their ranges so as to aid the search for AICP values in the bivariate $w^+ - Sb$ variegator. In all simulations of our model, it gave marginal bivariate coverage probabilities that were lower than univariate coverage probabilities, in qualitative agreement with the data in Lloyd *et al.* (1996). After matching marginal probabilities from data obtained for the bivariate case, certain AICP's were set to zero. This permitted us to obtain predictions for the two univariate cases.

The prediction for the w^+univariate variegator was in striking accord with the data. The prediction for the *Sb* univariate variegator was not a perfect match, but not wildly wrong. The overestimation of the probability of coverage for the *Sb* gene in the univariate case suggested a change in attractiveness to this position in the genome. We have verified that a slight change in the AICP's again matches the data exactly. This is a simple model that only uses cut-off values for gene coverage, quantities of molecules available, and relative attractiveness of the variegating breakpoint. Consequently, it only models competition between sites as dictated by the quantities of molecules and relative attractiveness of sites.

When affinities do not change in the presence of other chemical species, i.e., competition between sites is the only factor guiding chromatin assembly, our model confirms that the bivariate case can be used to generate the univariate case as seen for the w^{+} variegator.

In cases where the AICP's are not modified apart from setting the appropriate entries to zero as in Tables 2 and 3, the bivariate coverage of a gene is less than its univariate coverage reflecting competition. In other cases, such as the *Sb* gene, our model suggests that there is more to model than a competition effect. Specifically, our model suggests there is a change in attractiveness of a region of the genome as reflected in the AICP's we have used. In these cases, the bivariate case generates univariate coverage probabilities that are too high. This was seen to occur in the *Sb* variegator. Such more complicated cases may be typified by a bivariate coverage that is not sufficiently lower than the corresponding univariate coverage. This is the case for male flies in a double variegator context using *Sb* and the w^{m51b} allele of w^{+}: the coverage probability for *Sb* as a univariate variegator is the same as its coverage probability as a component in a double variegator (Lloyd, 1995; Lloyd *et al.*, 1996). We did not use this data set because the standard errors of the biological data were too large. In this case, the change in affinity or attractiveness would point to a factor that has not been adequately incorporated into the present model. There are many such factors, however, we are unable to conjecture which might be the most important. Moreover, we caution that in the present case, there is enough overparametrisation that it may be possible to modify the type 3 AICP's so as to match both the univariate cases and the bivariate case essentially exactly without assuming a change in attractiveness. We have resisted doing this on the grounds that it would be artificial as discussed in Section 3.

Our model formalises the notion that different sites compete for the available protein complexes. It distinguishes two cases on the basis of whether or not there is an affinity change in passing from the bivariate to a univariate case. In the absence of an affinity change, our model permits a prediction to be made from the bivariate case about the coverage probability of the univariate case. Since such affinity changes are necessarily small, our model gives approximate answers in that case. In addition, the model permits predictions to be made from the univariate experiments about the bivariate experiments. It generalises the simple case in which there are no overlapping complexes (here the type 3's) which follows easily from the rules of probability for independent random variables.

It is straightforward to perform the calculations described here so as to match other empirical probabilities given in Lloyd *et al.* (1996). However, we have not done this for two reasons. One, given any set of univariate or bivariate probabilities one can back-construct to find appropriate AICP's since there are too many degrees of freedom. Such matching is always possible mathematically and does not constitute a test of the model. Second, if one were to proceed with such matching in two related cases, it would be impossible to interpret the results. For instance, in the $w^{+} - Sb$ case we found $\xi = 38.5$ was "right". If, for instance a different reporter gene were used in place of *Sb*, one would probably get a different value for ξ. One would, however, be unable to make any reliable inferences as to what this would mean. If the new ξ were higher it could mean that *else* was more attractive or it could mean that the new species of complex for the new reporter gene behaved differently.

Despite these problems with existing data sets, there are various ways the present model can be subjected to further testing. First, recall that only estimates for the marginal probabilities were available. It is possible that redoing the experiment would permit one to obtain estimates for all four probabilities identified in Section 2.1. However, this may be impractical given present genetic techniques. Indeed, this could only be done if both reporter genes could be examined in the same cell. This is likely to be impossible for an eye color and bristle gene but is certainly possible for other variegating genes and would be a powerful test when those genes have been cloned or when antibodies to their protein products become available. This is likely within a few years given the efforts of the *Drosophila* Genome Project. Nonetheless, extending the matching to all four probabilities in the bivariate case so as to generate predictions for the univariate case would test the present methods. Second, it may be possible to impose further constraints on the model so as to reduce the number of free parameters. This was mentioned at the end of Section 3.3. For instance, we can require that the probabilities of type 3's going to *else* are constant when the first two entries are set to zero, and similarly for the corresponding boundary conditions in the other univariate case. We have not done this here because the differences in the type 3 AICP probabilities of going to *else* were small. Moreover, it is not clear that this, or any other physically plausible reduction in number of free parameters would be enough to get unique matching.

Further testing of this modeling approach is suggested by Table 2 in Lloyd *et al.* (1996). There, other alleles and reporter genes have been used to obtain estimates of coverage probabilities. In some cases, such as with the two alleles used here, sex differences have been noted. In particular, for female flies the bivariate coverage probabilities are the same as for the males whereas the univariate coverage probabilities both change. For males they were 0.91 for w^+ and 0.51 for *Sb*. For females they were both lower, 0.84 and 0.41 respectively. It is not clear what part of the present model makes it appropriate for male flies but not for female flies, however, the size of *else* may be one factor i.e., the relative proportion of heterochromatin in the genome.

In addition, the model may be tested by using other reporter genes, such as *brown* and other alleles of *white*. These would necessitate considering other protein complexes, possibly more than three, and having a cut-off value for the accumulation of each type at a site. In such cases, it might be necessary to have upper and lower cut-off values: A certain minimal number of complexes must bind to inactivate the gene and more than a certain number of complexes of each type cannot be bound. This neglects the possibility that the stoichiometry of the proteins found in different domains could differ. However, such an alternative hypothesis can be incorporated into the present model by increasing the number of species of protein complexes.

We anticipate that our modeling strategy will help to understand the distribution of non-histone chromosomal proteins. This is possible because a model such as this might permit identification of protein complexes that are common to a set of chromatin domains on the basis of the way they compete. From knowledge of where the breakpoints occur and the degree of competition for components one can in principle make inferences about which complexes typically bind to which locations. To see this consider the following series of experiments.

The point is to test if the protein complexes we have labeled type 3 are common to all variegating genes. Assume that they are common. Experiment 1 is the $w^+ - Sb$ double variegator examined here or an analogous experiment. Redo this experiment so as to obtain all four cell probabilities for the bivariate case. This is essential to remedy the overparametrisation problem. Experiment 2 is the same, but for the $w^+ -$ *brown* double variegator and experiment 3 is the *Sb* $-$ *brown* double variegator, or analogous experiments similarly compatible. Also, obtain the probability of coverage of *brown* as a single variegator. This gives a total of 15 data points—three univariate coverage probabilities and 12 bivariate coverage probabilities—four for each of the three pairs of alleles.

The first four rows of the AICP table has two free parameters, as does the second set of four rows. The last 16 rows have 12 free parameters. If this can be reduced, even by a couple of degrees of freedom, then there will be at least as many data points as free parameters. Consequently, from the first experiment we get one set of AICP's and from the second we get another. This second set has the same first four rows as the first set since w^+ is used. The second four rows correspond to a different type 2 complex, appropriate for *brown*. Now from these two experiments we have the first eight rows for the AICP's of the third experiment. The type 3's, which provide the competition, permit inferences to be made. If the protein complex species represented by type 3's is the same over all three experiments then it may be possible to identify elements in the type 3 AICP's over all three experiments, thereby permitting them to be estimated uniquely. (Note that if all can be identified there would be 18 degrees of freedom which might be reduced somewhat as described in Section 3.) If this cannot be achieved then one might infer that there are at least two species of protein complex that provide the competition.

This set of experiments would test the hypothesis that non-histone chromosomal proteins, constituents of complex types 1 and 2, are distributed in the genome in a non-random manner. That is, they are frequently associated with some chromatin domains and infrequently with others, in contrast to the histones or other proteins which might be more uniformly distributed.

Moreover, if two variegating genes must bind a certain number of shared components it is possible that one could outcompete the other so that both would be affected; one would be enhanced the other suppressed. One way to model this is to assume that a particular gene does not have a species of protein complex devoted specifically to it. In the $w^+ - Sb$ case for instance, we could model this by omitting type 2's completely and derive the coverage of *Sb* entirely from the shared type 3 components.

The consequence of this modeling approach is that diverse data sets might be synthesised into a coherent testable theory for the determination of number of protein complex species, their relative concentrations, and their approximate distribution over the genome.

4.2. REVERSE GRAVITY

A further complication to the model, qualitatively different from the modifications described in the last subsection arises from heterochromatin nucleation sites. Some researchers have postulated that chromatin starts to be assembled at a chromatin nucleation site on the DNA. Once the appropriate initial proteins have bound to the nucleation site, further assembly of chromatin is a semi-self-assembly process in that the molecules bound determine the molecules that will bind. This assembly process would provide a phenomenon that might be called "reverse gravity". The idea is that there should be a force which prevents the excess build-up of proteins at one site. That is, the distribution of the number of molecules that affix at a site should be limited but not tightly regulated. The accumulation of molecules at a site should by their very number discourage further molecules from binding until more have accumulated at other sites. This may arise due to proximity of a binding site to a heterochromatin nucleation site. The further a site is from a heterochromatin nucleation site as a result of having bound complexes the less it should attract further complexes from binding. Because of the reverse gravity, accumulations of complexes might tend to decrease the binding affinity of some sites relative to others.

Although we have not had to use this notion for the data matching we have undertaken, it can be incorporated into our model by including site dependent penalty factors which are increasing functions of the number of molecules which have already been bound at A_1, A_2, B_1 and B_2. Reverse gravity would provide increased stability to the probability that a gene is repressed. Operationally, the reverse gravity reweights the AICP's by factors which depend on the number of molecules that site has bound. Entries of the $\mathbf{X}(t)$ vector therefore appear in the reverse gravity expression. This introduces extra parameters that would have to be estimated.

REFERENCES

Boyes, J. & Bird, A. (1991). DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein. *Cell* **64,** 1123–1134.

Boyes, J. & Bird, A. (1992). Repression of genes by DNA methylation depends on CpG density and promoter strength: evidence for involvement of a methyl-CpG binding protein. *EMBO J* **11,** 327–333.

Elgin, S. C. R. (1995). *Chromatin Structure and Gene Expression*. Oxford: Oxford University Press.

Giese, K., Cox, J. & Grosschedl, R. (1992). The HMG domain of the lymphoid enhancing factor 1 bends DNA and facilitates assembly of functional nucleoprotein structures. *Cell* **69,** 185–195.

Grigliatti, T. A. (1991). Position-erect variegation: An assay for non-histone chromosomal proteins and chromatin assembly and modifying factors. In: *Functional Organization of the Nucleus* (Hamkulo, B. & Elgin, S. eds). pp. 587–627. San Diego CA: Academic Press.

Hayashi, S., Ruddell, A., Sinclair, D. & Grigliatti, T. (1990). Chromosomal structure is altered by mutations that suppress or enhance position-erect variegation. *Chromosome*, **99,** 391–400.

Heitz, E. (1934). Uber alpha-heterochromatin sowie konstanz und bau der chromomeren bei *Drosophila*. *Biol. Zentralbl.* **45,** 588–609.

Henikoff, S. (1981). Position-erect variegation and chromosome structure of a heat shock puff in *Drosophila*. *Chromosome* **83,** 381–393.

Hochstrasse, M., Mathog, D., Gruenbaum, Y., Saumweber H. & Sedat, J. (1986). Spatial organization of chromosome in the salivary gland nuclei of *Drosophila malanogaster*. *J. Cell Biol.* **102,** 112–123.

Lloyd, V. (1995). Genetic and Molecular Analysis of the *garnet* Eye Colour Gene of *Drosophila melanogaster*. Ph.D. Thesis, Department of Zoology, University of British Columbia.

Lloyd, V. K., Sinclair, D. A. & Grigliatti, T. A. (1996). Competition between different variegating rear rangements for limiting heterochromatic factors. *Genetics*, (submitted).

Locke, J., Kotarski, M. A. & Tartof, K. D. (1988). Dosage-Dependent modifiers of position erect variegation in *Drosophila* and a mass action model that explains their erect. *Genetics*, **120,** 181–198.

Paull, T. T., Haykinson, M. J. & Hohnson, R. C. (1993). The nonspecific DNA–binding and -bending proteins HMG1 and HMG2 promote the assembly of complex nucleoprotein structures. *Genes and Develop*. **7,** 1521–1534.

Sinclair, D. A. R., Mottus, R. C. & Grigliatti, T. A. (1983). Genes which suppress position-erect variegation in *Drosophila* are clustered. *Mol. Gen. Genet*. **191,** 326–333.

Wolff, A. P. (1994). *Regulation of Chromatin Strncture and Function*. Molecular Biology Intelligence Unit. R. G. Landes Co., Austin, Texas.

Wustmann, G., Szidonya, J., Taubert, H. & Reuter, G. (1988). The genetics of position-erect variegation loci in *Drosophila melanogaster*. *Nature*, **337,** 468–471.

Zucherkendl, E. (1974). Récherches sur les propriétes et l'activité biologique de la chromatine. *Biochimie*, **56,** 937–954.

APPENDIX

Mathematical Expressions for the AICP's

At time t, the number of type 1 complexes that have been used is $X_1 + X_3 + X_9$ and the number of type 1 complexes remaining is $P_1 - X_1 + X_3 + X_9$; the number of type two complexes that have been used is $X_5 + x_7 + X_{10}$ and the number remaining is $P_2 - X_5 + x_7 + X_{10}$; and the number of type 3 protein complexes that have been used is $X_2 + X_4 + X_6 + X_8 + X_{11}$ so $P_3 - X_2 + X_4 + X_6 + X_8 + X_{11}$ type 3 complexes remain. Thus at time $t + 1$, the probability of choosing a type 1 complex is

$$\frac{P_1 - (X_1 + X_3 + X_9),}{P_1 + P_2 + P_3 - t};$$

the probability of choosing a type 2 complex is

$$\frac{\text{[itP}_2 - (X_5 + X_7 + X_{10})}{P_1 + P_2 + P_3 - t};$$

and the probability of choosing a type 3 complex is

$$\frac{P_3 - (X_2 + X_4 + X_6 + X_8 + X_{11})}{P_1 + P_2 + P_3 - t}.$$

Once a complex of a type has been chosen, it must be sent to one of its possible sites. Thus, we condition on the type so as to define the probabilities of its binding to its permitted sites. This leads to the AICP's.

Denoting random variables by capitals and the corresponding outcomes by lower case, suppose the state of the system is summarised by the vector $(x_1, \ldots, x_{11}; c_1, \ldots, c_4)$ at time t and that complexes are chosen according to the above probabilities. If a type 1 complex is chosen then it must go to one of three sites: A_1, A_2, or *else*. Thus, at time $t + 1$, there are three states to which the system can go. The complex might go to A_1 so the state vector becomes $(x_1 + 1, x_2, \ldots, x_{11}; 0, c_2, c_3, c_3)$. The complex might go to A_2 so the state vector becomes $(x_1, x_2, x_3 + 1, \ldots, x_{11}; c_1, 0, c_3, c_4)$. Or, finally, the complex might go to *else* and the state vector becomes $(x_1, \ldots, x_9 + 1, x_{10}, x_{11}; c_1, \ldots, c_4)$. In this case the boundary does not change. Each of these possible states has a probability associated to it. The allocation of complexes to sites after having been selected at random from a pool makes this model hierarchical.

The probability of going from $(x_1, \ldots, x_{11}; c_1, \ldots, c_4)$ at time t to the state $(x_1 + 1, x_2, \ldots, x_{11}; 0, c_2, c_3, c_3)$ at time $t + 1$ can be written as the conditional probability

$$P((0, c_2, c_3, c_4, x_1 + 1, x_2, \ldots, x_{11}) \,|\, (c_1, \ldots, c_4, x_1, \ldots, x_{11})) = \frac{P_1 - (x_1 + x_3 + x_9)}{P_1 + P_2 + P_3 - t} P(\text{a type 1 goes to } A_1 | (c_1, \ldots, c_4)).$$

The probability of going from $(x_1, \ldots, x_{11}; c_1, \ldots, c_4)$ at time t to the state $(x_1, x_2, x_3 + 1, \ldots, x_{11}; c_1, 0, c_3, c_4)$ at time $t + 1$ can be written as

$$P((x_1, x_2, x_3 + 1, \ldots, x_{11}; c_1, 0, c_3, c_4) \,|\, (x_1, \ldots, x_{11}; c_1, \ldots, c_4)) = \frac{P_1 - (x_1 + x_3 + x_9)}{P_1 + P_2 + P_3 - t} P(\text{a type 1 goes to } A_2 | (c_1, \ldots, c_4)).$$

The probability of going from $(x_1, \ldots, x_{11}; c_1, \ldots, c_4)$ at time t to the state $(x_1, \ldots, x_9 + 1, x_{10}, x_{11}; c_1, \ldots, c_4)$ at time $t + 1$ can be written as

$$P((x_1, \ldots, x_9 + 1, x_{10}, x_{11}; c_1, \ldots, c_4) \,|\, (x_1, \ldots, x_{11}; c_1, \ldots, c_4)) = \frac{P_1 - (x_1 + x_3 + x_9)}{P_1 + P_2 + P_3 - t} P(\text{a type 1 goes to } \textit{else} | (c_1, \ldots, c_4)).$$

Note that $P(\text{a type 1 goes to } A_1 | (c_1, \ldots, c_4))$, $P(\text{a type 1 goes to } A_2 | (c_1, \ldots, c_4))$ and $P(\text{a type 1 goes to } \textit{else} | (c_1, \ldots, c_4))$ sum to one. They are the values of the AICP for the binding of a type 1 complex given boundary $(c_1,. \ldots, c_4)$. Note that the sum of the three time dependent probabilities gives the proportion of complexes remaining that are of type 1. The AICP's are not affected by time, by the complexes remaining or by the complexes bound, except for the boundary conditions one time step back. Similar reasoning applies if a type 2 or 3 protein complex is chosen at time t.

For any fixed set of AICP's, initial amounts P_1, P_2, P_3, and cutoff values M_1 and M_2, the protein complexes can be allocated and the Markov chain simulated so as to obtain estimates of the probabilities of the four events of interest, or their marginal probabilities. That is, we can find the probabilities of the events that the reporter genes are covered or not. These are the probabilities of the events

$$\{X_3 + X_4 \geqslant M_1, X_7 + X_8 \geqslant M_2\},$$
$$\{X_3 + X_4 \geqslant M_1, X_7 + X_8 < M_2\},$$
$$\{X_3 + X_4 < M_1, X_7 + X_8 \geqslant M_2\},$$
$$\{X_3 + X_4 < M_1, X_7 + X_8 < M_2\},$$

after $P_1 + P_2 + P_3$ steps. Each iteration gives a zero or one for each of these events. Getting a zero means the event did not occur; getting a one means the event did occur. Summing the sequence of zeroes and ones for each event and dividing by the number of iterations estimates the probability of the event.

For each boundary condition and each type to be bound there will be an AICP. As a result, it is seen that there are 24 AICP's in total, see Table 1 for a listing. There are four AICP's for type 1, corresponding to the boundary conditions of the form $(0, 0, c_3, c_4)$, $(0, 1, c_3, c_4)$, $(1, 0, c_3, c_4)$, and $(1, 1, c_3, c_4)$; the values of c_3, c_4 do not affect the binding of type 1 protein complexes. Likewise, since the boundary conditions c_1, c_2 do not affect the binding of type 2 complexes, there are $2^2 = 4$ AICP's for type 2. However, there are $2^4 = 16$ AICP's for type 3 since all four boundaries must be considered. For a given boundary condition for the binding of a type 1 complex, the AICP will consist of three probabilities—one for each site—which sum to one, as noted above. Similarly, for a given boundary condition, the AICP's for type 2 have 3 values which sum to one, since type 2 may also bind to three sites. The AICP's for type 3 have five entries that sum to one since type 3 complexes can bind to five sites.