Biostatistics (2002), **3**, 4, *pp.* 529–546
Printed in Great Britain

A Bayesian analysis of colonic crypt structure and coordinated response to carcinogen exposure incorporating missing crypts

JEFFREY S. MORRIS*
Department of Biostatistics, University of Texas, M.D. Anderson Cancer Center, 1515 Holcombe, Boulevard, Box 447, Houston, TX 77030-4009, USA
jeffmo@mdanderson.org

NAISYIN WANG
Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA

JOANNE R. LUPTON, ROBERT S. CHAPKIN, NANCY D. TURNER, MEEYOUNG HONG
Faculty of Nutrition, Texas A&M University, College Station, TX 77843-2471, USA

RAYMOND J. CARROLL
Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA

SUMMARY

This paper is concerned with modeling the architecture of colonic crypts and the implications of this modeling for understanding possible coordinated response of carcinogen–induced DNA damage between various regions of the colon. The methods we develop to address these two issues are applied to a particular important example in colon carcinogenesis. We cast the problem as an unusual and not previously studied hierarchical mixed-effects model characterized by completely missing covariates in units at a structurally base level, except for some randomly selected units. Information concerning the missing covariates is available through certain known ordering constraints and surrogate measures. Our methods use Bayesian machinery. We exploit the biological structure of this problem to generate the missing covariates simultaneously and efficiently at the base levels, as opposed to the naive practice of generating units at the base levels one-at-a-time with Metropolis–Hastings steps. We apply our methods to show that different regions of the colon have different architectures, and to estimate an important but non-standard function that measures the interrelationship of DNA damage mechanisms in different regions of the colon.

Keywords: Bayesian inference; Carcinogenesis; Colon cancer; Correlation; Functional data analysis; Gibbs sampler; Markov chain Monte Carlo; Missing covariates; Nutrition; Surrogate variables.

1. INTRODUCTION

This paper concerns the analysis of colon carcinogenesis data using a biologically-motivated hierarchical mixed-effects model with a partially missing covariate. Measurements of carcinogen-induced

*To whom correspondence should be addressed

DNA damage are obtained for individual colon cells from experimentally treated animals. The cells of the colon reside in clusters known as crypts.

We are interested in studying *structure*, or cell patterns within the crypts, and in modeling biologically processes underlying colon carcinogenesis in the presence of structure, when the structural covariate of interest is only available for randomly selected crypts. Our general approach can be used more broadly, but here we focus the investigation of possible *coordinated response* between the distal (back) and proximal (front) regions of the colon after carcinogen exposure resulting in DNA damage. Traditionally, statisticians have not been involved in the analysis of structure and coordinated response, so our modeling and computational efforts in this area represent new additions to the biological literature.

We use Bayesian machinery in our methods. To the best of our knowledge, no one has studied missing data problems with the structure we have encountered here. The models that we develop are thus the first that handle the type of missing data in our experiments, and they have the potential to serve as guides for other modeling efforts for missing data under ordering constraints. We also add that even if there were no missing data, the Bayesian formulation of our model has not been considered previously.

From the computational perspective, we introduce a new approach for handling the missing data. The naive, brute-force approach in our problem is to generate data 'one-cell-at-a-time', using Metropolis–Hastings steps, subject to the ordering constraints. This is not only slow and unwieldy, but also unnecessary. We develop what is in retrospect a simple efficient method of sampling the missing covariate data that exploits the basic biology inherent in the problem.

We now give some more details of the biology behind our problem, and explain what we mean by structure and coordinated response. The biochemical processes leading to tumor formation in the colon are not well understood. In investigating these processes, it is important to consider the special architecture of colonic epithelial cells. These cells spend their entire life cycles within surfaces of the colon wall that are known as crypts. The position a cell occupies within the crypt is related to its age, since new cells are formed in the stem cell region and move monotonically along the crypt wall as they mature. Figure 1 contains a cross-sectional image of some representative crypts from a rat's distal colon. The globular features lining the crypt walls are epithelial cells.

Since cells at different depths within the crypt are at different stages of maturity, they could by their very nature react differently to biological stimuli. Also, cells may have different levels of environmental exposure to carcinogens or other stimuli, depending on their relative distance from the lumen. Thus, it is crucial to consider some covariate indicating relative cell position within the crypt when modeling these processes in order to avoid missing any depth-specific effects. One important recent paper illustrating the crucial and often surprising role of cell position in colon carcinogenesis is that of Shih *et al.* (2001), which demonstrates that damaged cells at the top of crypts have the potential to form polyps, and later colon tumors.

1.1 *Structure and its importance*

It is typical in the colon carcinogenesis literature when measuring cell position to effectively assume that the cells are equispaced along the crypt wall (see for example Chang *et al.* (1997)). We have inspected many images of distal and proximal crypts, and it appears that the distal crypts clearly violate the equispaced assumption, and demonstrate particular patterns in the size and spacing of the cells, which we term *structure*. In Section 2, we investigate this issue in detail.

Why is structure, as we have defined it, important, in and of itself? In one sense, this is a 'because it's there' issue: at this time it is not known whether the cells are equispaced in the distal and proximal parts of the colon. If the distal and proximal regions have different cell position structure, then this could provide at least a partial explanation for the observed phenomena that the proximal and distal regions of the colon have different responses to environmentally induced exposures to carcinogens (Hong *et al.*, 2000).

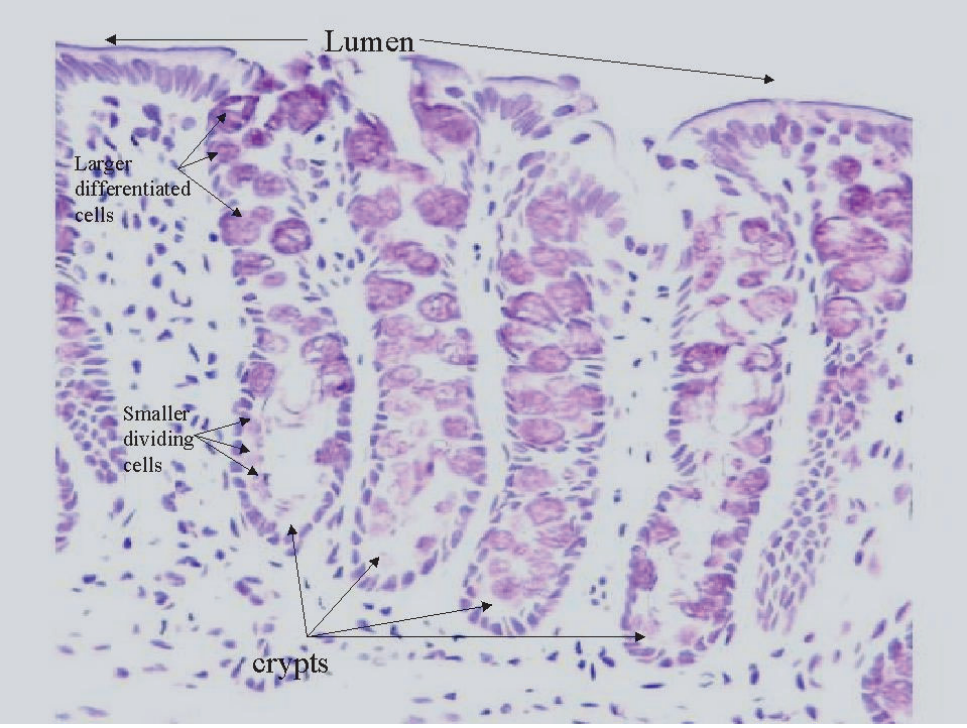

Fig. 1. Cross-sectional image of some representative crypts from a rat's distal colon.

An additional reason is as follows. The space a cell inhabits is a function of whether it is proliferating (smaller cells), differentiated (larger cells), or apoptotic (many different sizes). Different structures of the colonic crypts for the proximal and distal regions would thus suggest that their cell functions are not identical. Because of the monotonic progression of cells up the crypt as they mature, it is likely that more cells will be found within the same distance of the proliferating zone than are found outside that portion of the crypt. Inherent differences in the location of stem cells within the proximal and distal colon would suggest that an analysis of physiological function made assuming equispaced cells could lead to errors in interpretation. These principles are not restricted to the colon, but could be applied to the small intestine, as well.

1.2 *DNA damage in colon carcinogenesis*

Exposure to the procarcinogen azoxymethane (AOM) leads to DNA damage in the colon cells of rats, measured as *DNA adduct level*. This damage can eventually result in tumors if the damaged cells are not repaired or removed (Rogers & Pegg, 1977). The AOM-induced colon cancer in rats model is an important tool for delineating the mechanisms behind colon carcinogenesis. AOM must undergo a complex series of metabolic steps before being transformed to the ultimate carcinogen (Fiala, 1977), which is delivered to the colon both through the bloodstream and the lumen of the intestines. DNA adduct level, being an early biomarker for possible tumor development, is an important response to study in the initiation stage of colon carcinogenesis.

1.3 *Coordinated response and its importance*

The human body is a complex biological system with many interrelated parts, so there are limits in what can be learned by studying parts of the body separately. Important further information can be garnered by studying the possible interrelationships between processes operating across different organs, or different regions of the same organ. We refer to these interrelationships as *coordinated response*. These associations could be positive or negative, and could be due to any of a number of biological or environmental factors. The key idea is that by studying multiple organs or regions simultaneously, their interrelationships can yield additional insights not available from marginal analyses. There is much that is unknown about colon carcinogenesis, and investigation of coordinated response can yield some new insights into the process. There are indications in the colon carcinogenesis literature that carcinogenic behavior in one region of the colon can affect distant colonic regions (e.g. Barnes *et al.*, 1999).

In our particular example, we investigate coordinated response in the DNA damage process of two colonic regions. The question is whether DNA damage in the proximal region of the colon, after carcinogen exposure, is related to DNA damage in the distal part of the colon, i.e. the down-stream portion. No coordinated response between the regions would be an interesting finding in and of itself, because it would suggest that the different regions of the colon operate independently in responding to the carcinogen. Coordinated response, either positive or negative, would clearly also be of major biological interest, since it would suggest the mechanisms behind the activation and delivery of the carcinogen to colon cells in the two regions are linked. Finally, whether the coordinated response, or lack of coordinated response, depends on diet is of importance in understanding how diet can protect against colon cancer.

It is important to consider cell position when investigating this coordinated response, given that cells at different depths have different levels of environmental exposure and are inherently different biologically. It is possible that any interrelationships that exist between the DNA damage processes in the distal and proximal colon may depend crucially upon relative cell position.

Our goal is therefore to combine the analysis of structure with the analysis of coordinated response, and in so doing provide novel scientific insights into the early initiation stage of colon carcinogenesis.

1.4 *The simultaneous analysis of coordinated response and structure*

Morris *et al.* (2001) have already addressed coordinated response *in the absence of structure*. That is, using a particular measure of coordinated response described in Section 3, they assumed that cells were equispaced along colonic crypt walls. We call this equispaced assumption *nominal* cell position. They found a marginal (in the sense of statistical significance) negative relationship between the proximal and distal regions for rats fed a diet rich in corn oil, i.e. rats who had more DNA damage to the epithelial cells in the proximal region of the colon tended to have less damage in the distal region. The DNA damage mechanisms for rats fed a diet rich in fish oil in the two regions appeared to be unrelated, except that there was a positive relationship seen between the regions near the bottom of the crypt. The difference between the diet groups was marginally statistically significant, and is one indication, among many others (Hong *et al.*, 2000) that different diets can affect the degree of DNA damage to cells after carcinogen exposure.

In this paper, we address the coordinated response and structure issues simultaneously by measuring the physical cell position in crypts, normalized as is common to the interval $[0, 1]$, with 0 indicating the bottom of the crypt and 1 indicating the top. If we could measure physical cell positions on all the crypts in the analysis, then the methods of Morris *et al.* (2001) would apply. In what follows, we will call physical cell position *geometric* cell position.

What makes our experiment different is that measuring geometric cell position is tedious and very time-consuming. Thus, the data we have obtained is that for each rat, on a small subset of colonic crypts,

both geometric and nominal cell positions are ascertained, while on the vast majority of crypts, only nominal cell positions are measured.

In statistical terms, we have a missing data problem that to the best of our knowledge has not been addressed in the literature. Specifically, the complete data are hierarchical, with observations on rats at the top level, then observations on multiple crypts within each rat at the next level, and finally observations on DNA damage, nominal call position and geometric cell position at the base level. What is unique about our data is that for most crypts, at the base level *all* observations of the interesting covariate, geometric cell position, are missing.

We address this problem using Bayesian machinery, although clearly likelihood methods could also be employed. As we will describe later, each colonic crypt contains on the order of 30 cells, the geometric positions of which form an ordered sequence in the unit interval. Any MCMC or EM computation necessarily must generate observations from the missing geometric cell positions. The naive, most straightforward and brute-force approach is to generate these cell positions 'one-cell-at-a-time' using a Metropolis–Hastings algorithm, while respecting the ordering. Our approach is different. We use our model for geometric cell position to suggest a simple and fast means of generating the 'missing data'. The method is much faster than the brute-force approach, and in retrospect it is also exceedingly simple.

Finally, we address an issue raised by most readers of a draft of this paper. Specifically, what is the 'best' scale to measure cell position, and hence to understand coordinated response of DNA damage and whether it depends on cell position? Our own view is that the geometric cell position is most intuitive and, in general, most appropriate. It contains information on both the age and stage of maturity of the cell as well as the relative distance from the lumen, while the nominal cell position only contains information on the former.

In the presence of structure, it is quite possible that using nominal cell position in place of geometric cell position can lead to misleading results, especially in cases where proximity to the lumen is of potential importance.

1.5 *Outline*

The paper is ordered as follows. Section 2 describes the structure analysis. Section 3 combines structure, coordinated response and missing data. Section 4 discusses some implications of our results and the general applicability of our method.

In a large sense, this paper was motivated by the negative relationship in corn-oil-fed rats that was described above. This result was surprising to all of us, because if there was any coordinated response, we had expected it to reveal a positive relationship. The geometric cell positions were thus measured with a view towards robustness: to see if the negative correlation was a by-product of using the nominal cell positions. Since geometric cell positions are time-consuming to measure, we expect that any further analysis of colon cancer data that uses them will confront the type of missing data that we have encountered, so our method will prove to be useful.

2. CELL POSITION: THE STRUCTURE OF COLONIC CRYPTS

In this section we investigate the structure of colonic crypts, specifically the distributions of the geometric cell positions within crypts from the distal and proximal colon. Section 2.1 describes the data structure, Section 2.2 describes the model, and Section 2.3 describes the results. In addition, Section 2.4 describes different modeling assumptions and their impact on the analysis.

2.1 *Data structure*

We define the nominal and geometric cell positions precisely as follows. By definition, for the nominal position, the bottom of the crypt is coded as relative cell position 0, and the top of the crypt is relative cell position 1. Given n cells lining one wall of a particular crypt, the nominal cell position for the ith cell from the bottom is $i/(n+1)$. When the physical location of each cell is known, the geometric relative cell position is defined to be the ratio of the actual distance from the bottom of the crypt to the middle of the cell, and the length of the crypt.

The main data set used in this paper consists of measurements taken from roughly 20 distal and proximal colon crypts per rat on three rats from each of ten treatment groups. These rats were fed one of two diets, supplemented with either fish oil or corn oil. The rats were then exposed to a colon carcinogen azoxymethane (AOM), and euthanized after a set period of time. The DNA adduct level was determined for each cell lining one wall from a cross-sectional cut of each selected crypt. A treatment structure of a 2×5 factorial with factors diet and time from exposure to the carcinogen until euthanization was considered, with DNA adduct level in a colonic crypt cell the response variable. The nominal cell positions were then computed for all crypts. More details about the experiment and the fundamental biology of crypt cell dynamics are given in Hong *et al.* (2000) and Morris *et al.* (2001).

A validation data set, consisting of the geometric cell positions for three randomly selected crypts from the distal and proximal colon for each rat, was also obtained.

2.2 *Model for Geometric cell positions*

The cell structure within colonic crypts imposes a natural ordering of the geometric cell positions within each crypt. The natural brute–force model would be of GEE-type, i.e. specify the marginal distributions for each cell, along with the necessary ordering constraints. While it is certainly possible to specify such a model, doing so is unwieldy compared to the alternative we suggest. We propose to model all the cells within the crypt jointly as the order statistics of a Beta(a, b) distribution. This modeling choice accounts for the inherent ordering constraints and naturally accommodates a flexible class of possible structures. For the sake of parsimony, we assume the same Beta distribution for all crypts within the proximal region, although see Section 2.4 below for more discussion. For the distal region, we use a different Beta distribution, but similarly assume a common distribution for all crypts within the region.

The distal and proximal Beta parameters are denoted by (a_d, b_d) and (a_p, b_p), respectively. Denote the number of cells lining the left wall of the distal crypt c_d and proximal crypt c_p for rat r of time group j for diet group i by n_{ijrc_d} and n_{ijrc_p}, respectively. The likelihood for (a_d, b_d) given the distal geometric cell positions X_{ijrc_dl}, $(i = 1, 2;\ j = 1, \ldots, 5; r = 1, \ldots, 3; c_d = 1, \ldots, 3; l = 1, \ldots, n_{ijrc_d})$ is

$$L(a_d, b_d) = \prod_{i,j,r,c_d,l} [\{B(a_d, b_d)\}^{-1} (X_{ijrc_dl})^{(a_d-1)} (1 - X_{ijrc_dl})^{(b_d-1)}], \tag{1}$$

where $B(a, b) = \{\Gamma(a)\Gamma(b)\}/\Gamma(a + b)$, with $\Gamma(\cdot)$ being the complete Gamma function. A similar likelihood holds for the proximal parameters given the proximal cell positions.

2.3 *Distribution of Geometric cell positions*

We examine the distribution of geometric cell positions by computing the posterior distributions of the parameters (a_p, b_p) and (a_d, b_d), respectively, given the validation data set. Independent Gamma priors are chosen for a_d, b_d, a_p, and b_p, where the Gamma(α, γ) density is defined to be $f(x|\alpha, \gamma) = \{\gamma^\alpha / \Gamma(\alpha)\} x^{\alpha-1} \exp(-\gamma x)$. The hyperparameters for a_d, namely $(\alpha_{ad}, \gamma_{ad})$, are chosen to have unit mean with a large variance: similar notation and prior distributions are used for the other hyperparameters.

Define $B(a, b)$ as above to be the beta-function with arguments (a, b). The log of the posterior density of a_d and b_d given the distal cell position data is proportional to

$$-\sum_{i,j,r,c_d} n_{ijrc_d} \log\{B(a_d, b_d)\} + (a_d - 1)\left[\sum_{i,j,r,c_d l} \log(X_{ijrc_d l})\right] \tag{2}$$
$$+(b_d - 1)\left[\sum_{ijrc_d l} \log(1 - X_{ijrc_d l})\right] - \gamma_{ad} a_d - \gamma_{bd} b_d + (\alpha_{ad} - 1)\log(a_d) + (\alpha_{bd} - 1)\log(b_d).$$

A similar equation holds for the proximal parameters.

Samples from the posterior distribution are obtained using a Metropolis–Hastings algorithm (Gilks *et al.*, 1996). Details necessary for implementation of the procedure can be found in Appendix A.

Sensitivity analyses show that this choice of vague proper priors had little impact on the results. The parameter σ^2_{ab} in the Metropolis–Hastings algorithm (see Appendix A) was set to be 0.001. With these settings, a new candidate value for the Metropolis–Hastings algorithm was accepted with probability 54% for distal and 60% for proximal regions. Multiple chains of the Metropolis-Hastings were obtained with various starting values, and all clustered around the same distributions. A convergence test (Gelman & Rubin, 1992) was conducted separately for each parameter and indicated that convergence was attained. The results reported here are from a single chain of 100 000 Metropolis-Hastings iterations after a burn-in time of 10 000.

Figure 2 contains the histograms demonstrating the posterior distributions of a_d, b_d, a_p, and b_p. The posterior means for the proximal parameters are $(0.94, 0.97)$, and the posterior probabilities of them being greater than one are 0.01 and 0.15, respectively. These results seem to suggest that for the proximal crypts, there is some evidence of nonuniformity in the distribution of the geometric cell positions. In spite of the statistical evidence of non-uniformity, little practical difference between the estimated proximal distribution and the uniform proximal distribution is evident upon comparison of the two.

For crypt cells from the distal colon, the parameters have posterior means of $(0.84, 1.01)$, and the posterior probabilities of them being greater than one are 0.00 and 0.83, respectively. These results suggest a systematic departure from uniformity in the geometric cell position distributions in crypts from the distal colon. Figure 3 contains plots of the geometric versus nominal cell positions for randomly selected distal crypts, which demonstrate the departure from uniformity. Note that cells in crypts from the distal colon are consistently concentrated at the bottom of the crypts.

Thus, our analysis suggests that the uniform model is acceptable for the geometric cell positions in crypts along the proximal colon, but not in the distal colon. In the distal colon, we observe a systematic concentration of crypt cells towards the bottom of the crypts. In other words, the proximal and distal regions differ in the structure of their colonic crypts.

Armed with this quantitative information that the two regions have crypts of different structure, we examined many crypts visually. In general, it appears to us that cells at the bottom of crypts from the distal colon tend to be smaller and arranged closer together relative to cells higher up the crypt. This difference in cell size and proximity is not generally visualized in crypts from the proximal colon. One can speculate as to why it is that the distal region seems to have a different cell structure than the proximal region, but at this point we have no compelling argument. This variability in cell size and location may be related to the age of the cell and to its stage of maturation and differentiation as it moves away from the stem cell of the crypt. Given that stem cells are known to be located at the bottom of the distal crypts, the small size and bunching of cells near the bottom of the crypt is expected to be due to the production of many immature cells that will migrate as they mature. This observed phenomenon is consistent enough from crypt to crypt to result in a distribution of cell positions across crypts that differs significantly from that of a uniform distribution. There is some uncertainty regarding the location of stem cells in crypts from the proximal

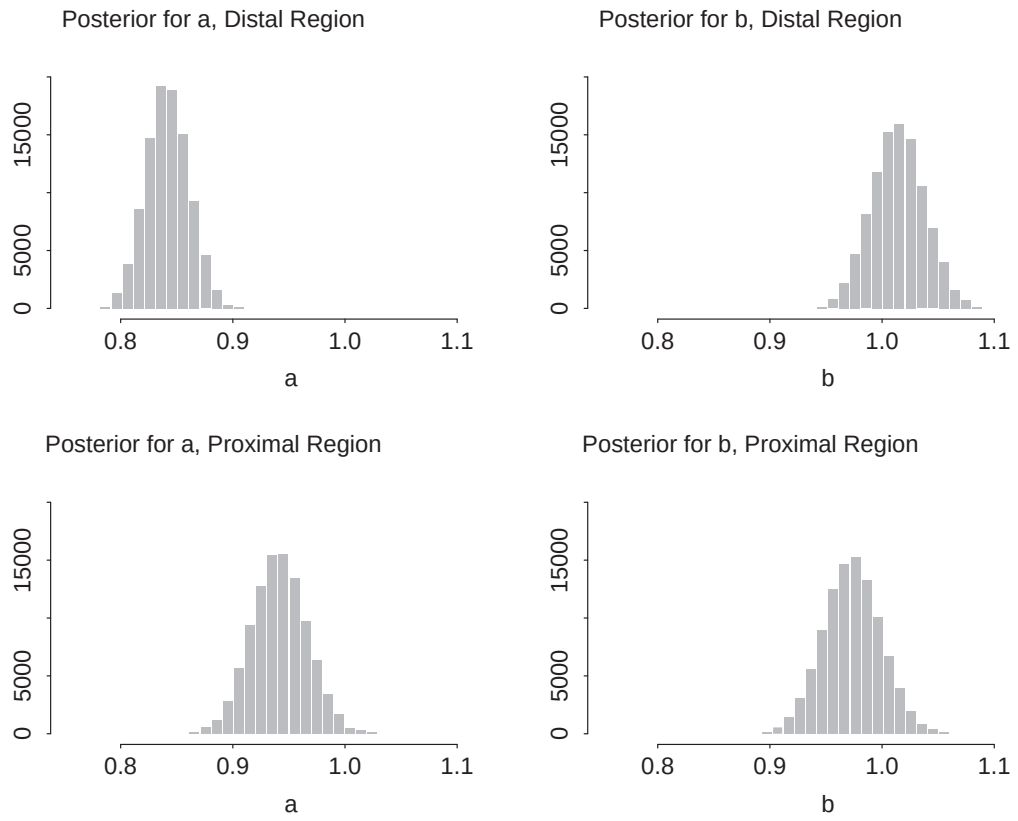

Fig. 2. Estimated posterior distributions of Beta hyperparameters in proximal and distal colon.

colon, although the bottom of the crypts has been considered as the likely location for these stem cells, as well. Sato & Ahnen (1992) have hypothesized that the stem cells are in fact located more towards the middle of proximal crypts, with cells migrating in both directions as they mature. The resulting suggestion from this hypothesis, which is that the cells in the middle of proximal crypts would be smaller and closer together, is not generally observed in our validation data set. The validation data set indicates variability in the crypts of the proximal colon in that the cells of some crypts are concentrated at the bottom, and others demonstrate cell concentration at the top, while still other crypts seem to exhibit a roughly equispaced distribution of cells.

2.4 *Possible alternative models and sensitivity analyses*

We focus here for specificity on the distal region of the colon. Our model assumes that the geometric cell positions in the crypt are distributed as the order statistics of a Beta(a_d, b_d), where (a_d, b_d) are the same for every time, diet, rat and crypt. We denote this by $\underline{X}_{trc}(d) = \text{o.s. Beta}(a_d, b_d)$.

The natural generalization of this model is to allow the Beta parameters to vary in some way over time periods, diets and rats, but to be the same across crypts within rats. In this early initiation phase, it makes no biological sense to think that the structure of the crypts will vary over diet and time, but it may make

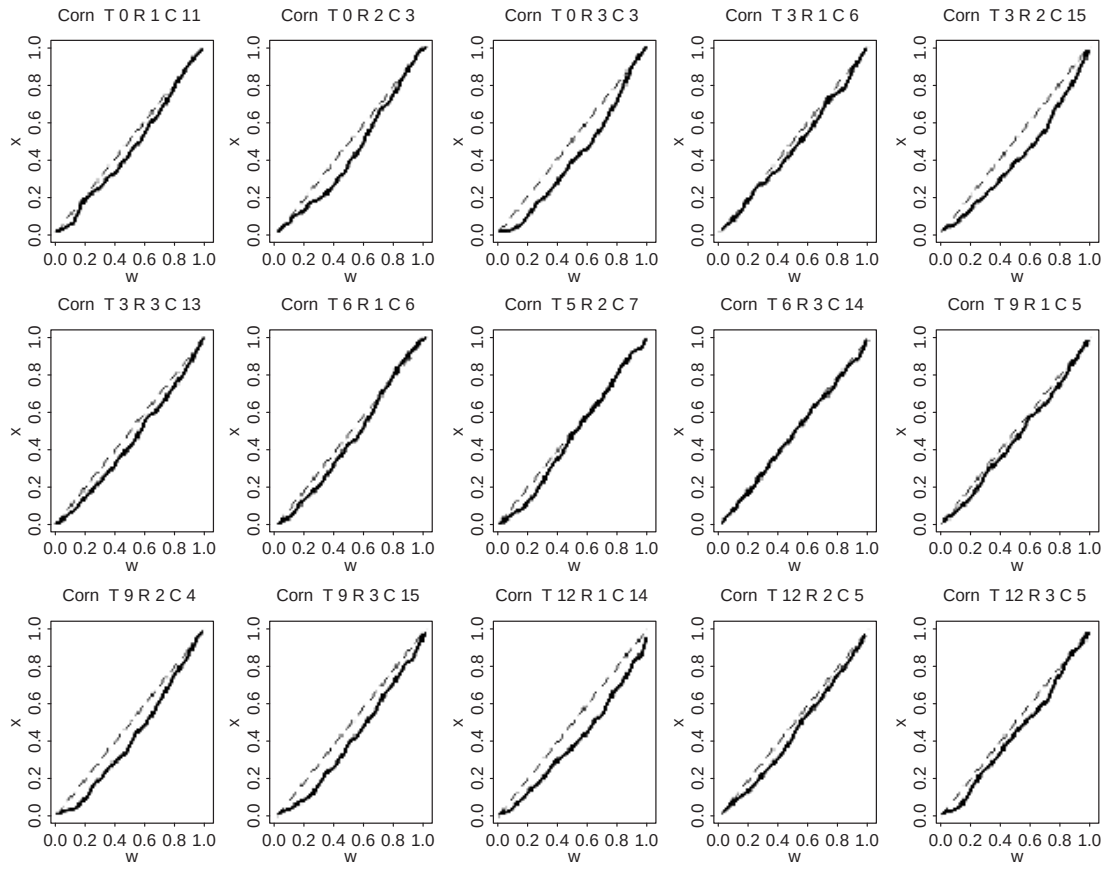

Fig. 3. Plots of geometric cell positions (X) versus nominal cell positions (W) for selected distal crypts in validation data set.

sense to allow the structure of crypts to depend on the rat. Thus, a natural hierarchical model is

$$\begin{aligned} \underline{X}_{trc}(d) &= \text{o.s. Beta}(a_{trd}, b_{trd}); \\ a_{trd} &= \text{Gamma}\{\text{mean} = \mu_a(d), \text{var} = \sigma_a^2(d)\}; \\ b_{trd} &= \text{Gamma}\{\text{mean} = \mu_b(d), \text{var} = \sigma_b^2(d)\}, \end{aligned} \tag{3}$$

with prior distributions $\mu_a(d) = \text{Gamma}(\mu = 1, \sigma^2 = \text{large})$, $\mu_b(d) = \text{Gamma}(\mu = 1, \sigma^2 = \text{large})$, $\sigma_a^2(d) = \text{IG}(\mu = 1, \sigma^2 = \text{large})$, $\sigma_b^2(d) = \text{IG}(\mu = 1, \sigma^2 = \text{large})$. The complication is that instead of having to develop Metropolis–Hastings steps only for (a_d, b_d), now we need steps for $a_{tr}(d)$, $b_{tr}(d)$, $\mu_a(d)$ $\sigma_a^2(d)$, $\mu_a(d)$ and $\sigma_a^2(d)$.

A further generalization is to allow the Beta parameters in (3) to depend on the crypt as well, and then build a further step into the hierarchy.

Instead of embarking on these extra calculations, we instead did a type of marginal analysis. Specifically, we treated the crypts as fixed, so that $\underline{X}_{trc}(d) = \text{o.s. Beta}(a_{trcd}, b_{trcd})$, where (a_{trcd}, b_{trcd}) are treated as separate parameters for every d, t, r and c, whose prior distributions are Gamma with mean one and a large variance. We fitted this model, and computed the posterior means of (a_{trcd}, b_{trcd}) given

the data. We then computed the means of these posterior means across times, diets, rats and crypts. The results were almost identical to the posterior means found in Section 2.3. The same near-identical results were found for the proximal region.

This model-sensitivity analysis gives additional support to our conclusion that the distal crypts have cells that are not equispaced.

3. CELL POSITION: COORDINATED RESPONSE BETWEEN REGIONS OF THE COLON

The purpose of this section is to discuss our analysis of coordinated response when combined with structure. In Section 3.1, we describe the function that we use to measure coordinated response. Section 3.2 gives some theory as to why misleading inferences about the correlation may be obtained if nominal and not geometric cell position is used. In Section 3.3 we define the model used in the analysis. Details of the Gibbs sampler are described in Section 3.4. The actual analysis of the data is discussed in Section 3.5. A sensitivity analysis under a different model is described in Section 3.6

3.1 *Correlation function as a measure of coordinated response*

While there are many ways to measure coordinated response between regions of the colon, a convenient means is through a type of correlation function originally developed in Morris *et al.* (2001). The conceptual basis for this correlation function is as follows. First, the number of actual colonic crypts in any region of the colon is enormous. Consider a given cell depth W, normalized as before to the unit interval. For any rat r, consider all the cells (conceptually nearly infinite in number) at depth W_1 in the proximal region, and compute the mean response $Y_r(p, W_1)$. Do the same thing in the distal region at depth W_2 to obtain $Y_r(d, W_2)$. Let $\rho(W_1, W_2)$ be the correlation across rats between $Y_r(p, W_1)$ and $Y_r(d, W_2)$. This is the correlation function of interest, with the major special case being when $W_1 = W_2$.

Quite clearly, the correlation function $\rho(W)$ at a common depth W measures a type of coordinated response, the importance of which has already been described in Section 1.3. Given the importance of cell position within the crypt, it is also of major interest to understand whether the coordinated response, if it exists, depends on cell position.

With cell position measured on the nominal scale, an analysis of coordinated response was performed by Morris *et al.* (2001). They concluded that the estimated correlation function for rats fed a diet supplemented with corn oil while exposed to AOM was negative at all depths of the crypt, and basically constant. This negative correlation was not expected, especially in light of a slightly positive correlation function near the nominal bottom of the crypt for rats fed a diet rich in fish oil. In effect then, they found that the damage process in the proximal part of the colon was linked with the damage process in the distal part of the colon in different ways depending on the animal's diet.

To us at least, nominal cell position, called W from here on, is an inherently less satisfactory measure than geometric cell position, called X. Given the results of Section 2, one would expect somewhat similar answers in either scale, because the proximal region cells are very nearly uniformly distributed, while the distal region cells, while distinctly non-uniform, are not ridiculously so. Thus, our main purpose in this section is to develop a method for estimating the correlation function for rats fed dietary corn oil, based on the geometric cell positions, X, and compare it with the function based on the nominal cell positions, as computed by Morris *et al.* (2001). Recall that the validation data set contains geometric cell position measurements for only a randomly selected portion of the crypts in the original AOM study. The nominal cell position, W, is available for all cells of all crypts and effectively serves as a surrogate measure in our method.

3.2 *Theoretical analysis*

In this section, we give a brief theoretical explanation as to why using nominal instead of geometric cell position may give misleading inferences.

Consider for theoretical purposes the situation that there are an infinity of cells at every geometric cell depth. Suppose that the distal cells are distributed as Beta(a_d, b_d) while the proximal cells are distributed as Uniform[0, 1]. If the distal geometric cell position is s and the proximal geometric cell position is t, and we average over the infinity of cells for any given rat, suppose that the rat-level observation is a simple linear random coefficient model:

$$Y_r(d, s) = \beta_{0r}(d) + \beta_{1r}(d)s; \tag{4}$$

$$Y_r(p, t) = \beta_{0r}(p) + \beta_{1r}(p)t. \tag{5}$$

In (4), (5), the random intercepts and slopes have zero mean, variance equal to one, and are mutually independent except that for some η,

$$\text{cov}\{\beta_{1r}(d), \beta_{0r}(p)\} = -\text{cov}\{\beta_{0r}(d), \beta_{1r}(p)\} = \eta \neq 0. \tag{6}$$

By construction, at the same cell depth, (4), (5) are uncorrelated. Indeed, $\text{cov}\{Y_r(d, s), Y_r(p, t)\} = \eta(s - t)$, and $\text{cov}\{Y_r(d, t), Y_r(p, t)\} = 0$.

We now show that if nominal and not geometric cell position were used for the crypts, then at the same *nominal* depths, a correlation would be found. Let $\text{Beta}^{-1}(t, a_d, b_d)$ be the inverse cdf of the Beta(a_d, b_d). If the nominal cell position for distal crypts is t, then the geometric cell position is $s(t) = \text{Beta}^{-1}(t, a_d, b_d)$, and hence in the nominal scale the correlation is

$$\text{corr}\{Y_r(d, \text{nominal} = t), Y_r(p, \text{nominal} = t)\} = \eta\{s(t) - t\}/[\{1 + s^2(t)\}(1 + t^2)]^{1/2}. \tag{7}$$

Figure 4 plots (7) when $\eta = 0.5$, for choices $a_d = 0.1, 0.3, 0.5, 0.8$ and $b_d = 2 - a_d$. As seen in that figure, even in this simple case there is the potential for concluding a negative correlation from nominal cell position data when there is no such correlation in geometric cell position data.

3.3 *The model*

The DNA adducts were modeled using a hierarchical structure. The model for the DNA adduct level for cell l of distal crypt c_d and proximal crypt c_p for rat r in time group j, fed a diet supplemented with corn oil, denoted Y_{jrc_dl} and Y_{jrc_pl}, with geometric relative cell positions X_{jrc_dl} and X_{jrc_pl}, respectively, is given by

$$\begin{aligned} Y_{jrc_pl} &= Z_{c_p}(X_{jrc_pl})\beta_{jrc_p,p} + \epsilon_{jrc_pl,p}; \\ Y_{jrc_dl} &= Z_{c_d}(X_{jrc_dl})\beta_{jrc_d,d} + \epsilon_{jrc_dl,d}, \end{aligned} \tag{8}$$

where $Z(X)$ is a design matrix that is a function of the geometric cell positions X, $\beta_{jrc_p,p}$ and $\beta_{jrc_d,d}$ are crypt-level coefficients and $\epsilon_{jrc_pl,p}$ and $\epsilon_{jrc_dl,d}$ are the residual errors for the proximal and distal measurements, respectively. The residuals are assumed to be independent between crypts, and for simplicity, also within crypts. As discussed in Morris *et al.* (2001), this condition effectively holds for our sample. The proposed approach can be modified to weaken this assumption when it is necessary. The error distribution is assumed to be Gaussian with mean zero and variances $\sigma^2_{\epsilon,d}$ and $\sigma^2_{\epsilon,p}$, respectively. We used the same model as in Morris *et al.* (2001), so that DNA damage was modeled as a quadratic function of cell position.

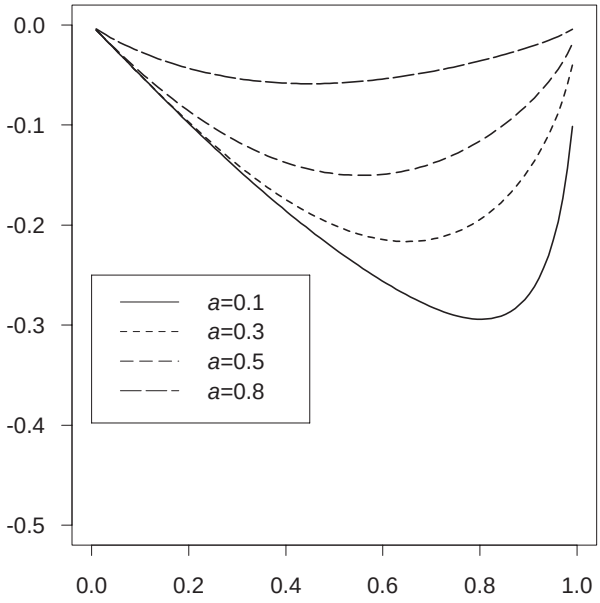

Fig. 4. Plots of the correlation function for the same cell depth as estimated by using nominal cell positions when the distal cell positions have a Beta distribution with arguments a and $b = 2 - a$. Here the correlation $\eta = 0.5$ in (6). The correlation function is identically zero when geometric cell positions are used.

Coefficients at the crypt level, $\beta_{jrc_p,p}$ and $\beta_{jrc_d,d}$ are assumed to be independent from a Gaussian distribution with means of coefficients at the rat level $\beta_{jr,p}$ and $\beta_{jr,d}$ and respective covariance matrices $\Sigma_{C,p}$ and $\Sigma_{C,d}$. Sample crypts originate from either the distal or proximal segments of the colon, therefore DNA adduct levels associated with a proximal colon cell sample or a distal colon cell sample cannot be elicited from a single crypt. This is the motivation for the independence assumption stated above. Coefficients at the rat level are assumed to follow a Gaussian distribution with means being the coefficients from the time level, $\beta_{j,p}$ and $\beta_{j,d}$, covariance matrices $\Sigma_{R,p}$ and $\Sigma_{R,d}$ with cross-covariance matrix $\Sigma_{R,pd}$. The combined rat-level covariance matrix is denoted Σ_R. Of primary interest for this study is the correlation function at the rat level, $\rho(X)$. This represents the relationship between the DNA adduct levels of crypt cells from the distal and proximal colon as a function of relative cell position X, given by

$$\rho(X) = \{Z_p'(X)\Sigma_{R,pd}Z_d(X)\}\{Z_p'(X)\Sigma_{R,p}Z_p(X)\}^{-1/2}\{Z_d'(X)\Sigma_{R,d}Z_d(X)\}^{-1/2}. \tag{9}$$

We observe the geometric cell positions X for each cell within the crypts randomly selected to be included in the validation data set, but these values are missing for all cells within the other crypts. The nominal cell position, W, is observed for all crypt cells.

We assume that the vectors $\underline{X}_{jrc_d}$ and $\underline{X}_{jrc_p}$ follow the distribution of the order statistics of a sample of size n_{jrc_d} and n_{jrc_p}, respectively, from a Beta distribution with parameters (a_d, b_d) and (a_p, b_p), just as in Section 2.2. Note that this specification implicitly relates the covariate X and surrogate W, since the distribution of X depends on n (the number of cells for a given crypt), and the vector of W for a crypt is determined entirely by n.

3.4 *Estimation approach*

An estimator of the correlation function can be constructed from (9), based on the estimated covariance at the rat level, Σ_R. This is done using a Bayesian framework by specifying priors for the unknown parameters at the lowest level of the hierarchy for the model described in Section 3.3. The priors are chosen in the customary way. Specifically, Inverse Gamma priors are chosen for the residual variances $\sigma^2_{\epsilon,d}$ and $\sigma^2_{\epsilon,p}$. The priors on the rat and crypt level covariance matrices are chosen to be of an Inverse Wishart form (Muirhead, 1982, p. 97); and independent Gaussian priors are chosen for the coefficients at the time levels $\beta_{j,p}$ and $\beta_{j,d}$. As in Section 2.3, independent Gamma priors are assumed for the Beta hyperparameters as well. For the coefficient and covariance parameters, we choose vague proper priors centered at empirical Bayes estimates with large variances. Sensitivity analyses indicate that these priors do not have a strong impact on the results. The prior for the rat-level covariance matrix was chosen so the mean prior correlation function was zero for all relative cell positions. Again, we use multiple chains with diverse starting values for the parameters, and the different chains converged to the same distributions.

The estimated posterior distribution of the correlation function is obtained using the posterior distribution of Σ_R based on (9), which is in turn estimated using a Gibbs sampler (Tanner, 1996). The complete conditional distributions for the various coefficient and covariance parameters can be found in closed form since conjugate priors were assumed. A Metropolis-Hastings step is used to update the Beta hyperparameters. The results from Gelman–Rubin tests performed on the individual parameters suggest convergence is obtained for the model parameters. The results presented here are from a single chain of 10 000 Gibbs samples obtained after a burn-in time of 1000. Details are given in Appendix B.

Thus, as advertised earlier, for the most part the Bayesian calculations proceed along familiar lines, although the details are new because the model is new. The main exception is the generation of the missing geometric cell positions, a feature of this problem distinct from the usual hierarchical modeling framework.

The brute-force method is to generate the geometric cell position one-cell-at-a-time. This is possible to do: for example, for an interior cell with nominal cell position W and neighboring geometric cell positions x_1 and x_2, one need generate a Beta random variable constrained to lie in the interval $[x_1, x_2]$. The cell-at-a-time method leads to very slow computation.

We use a different approach, which is far simpler and faster. All missing cell position measurements within a given missing crypt are generated **simultaneously** to ensure the ordering constraints on the missing cell positions are met. This is easily achieved with our modeling strategy, which specifies the joint distribution of cell positions within a crypt as the order statistics of a Beta distribution, even though their marginal distributions are difficult to specify. This strategy greatly simplifies the analytical details of the problem, and considerably reduces the computational complexity of the fitting of the model as well. Some details of the Metropolis-Hasting step used to generate these missing cell positions can be found in Appendix A.

We believe that our approach can be used in other situations that a covariate vector is missing entirely at the base level of analysis, when that covariate vector is constrained to lie in the unit interval and some surrogate is available.

3.5 *Application to AOM data*

The posterior mean correlation function for corn-oil-fed rats and 90% posterior bounds based on the geometric cell positions are given in Figure 5. The corresponding estimated correlation and bounds based on the nominal cell positions are also given for comparison.

For this data we see a close match between the correlation function based on the geometric and nominal cell positions. In spite of the fact that the distributions of cells within a crypt differ from the distal

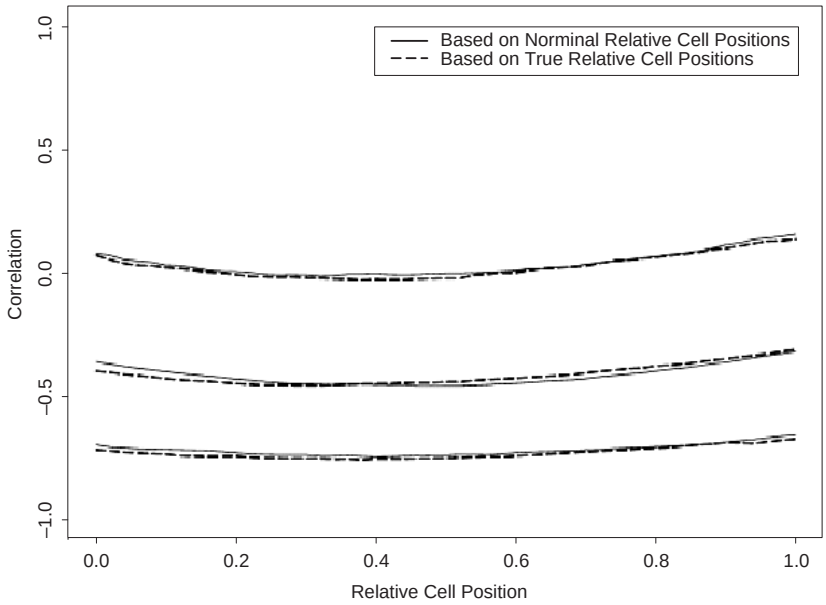

Fig. 5. Estimated correlation function between DNA adduct levels in crypt cells within proximal and distal regions as functions of nominal and geometric cell positions, along with 90% posterior bounds.

colon to the proximal colon, a negative correlation was still observed at all depths within the crypts for corn-oil-fed rats. Thus, we conclude that the estimated negative correlation was not an artifact associated with use of the nominal scale.

3.6 *Possible alternative models and sensitivity analyses*

As shown above, for the data the correlation functions in the nominal and geometric cell position scales are virtually identical. As shown in Section 3.2, the two correlation functions can in principle be quite different. Indeed, in our own data, had the cell positions been distinctly non-uniform, a similar phenomenon would have occurred.

In Figure 6 we plot the correlation functions for the nominal and geometric scale for rats fed a diet rich in fish oil. We also replaced the measured geometric cell positions by cell positions generated by two different Beta distributions: Beta$(0.1, 1.9)$ and Beta$(1.9, 0.1)$. In both cases, there are fairly large differences with the nominal/geometric correlation functions: in the former case the positive correlation disappears, while in the latter case, the near-zero correlation midway up the crypt becomes noticeably positive.

Thus theoretically, and now empirically, we have demonstrated that it is possible for the nominal and geometric scales to lead to visually different correlation functions.

4. CONCLUSIONS

We have discussed a biologically-motivated hierarchical mixed-effects model. The hierarchy involves animals, colonic crypts within animals, and cells within colonic crypts. A biologically meaningful covariate is geometric cell position, measured within crypts and taking on values in the unit interval.

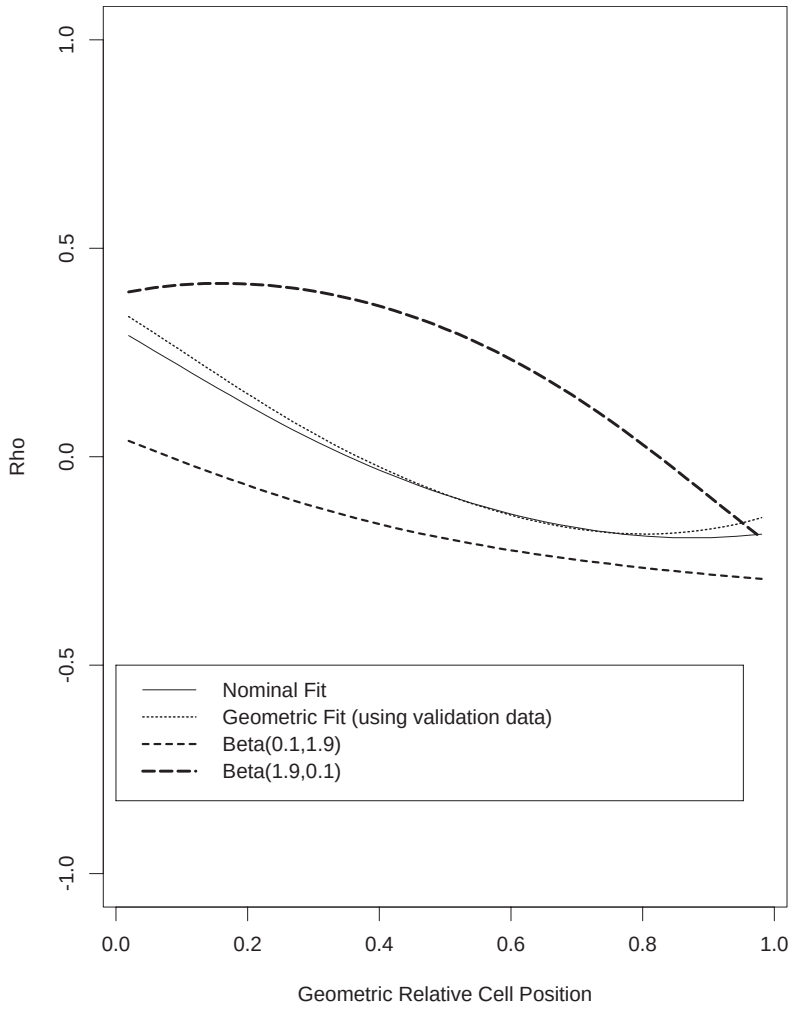

Fig. 6. Estimated correlation function between DNA adduct levels in crypt cells within proximal and distal regions as functions of nominal (solid), geometric (dotted), assuming that geometric cell positions follow a Beta(0.1, 1.9) (small dashed) and assuming that geometric cell positions follow a Beta(1.9, 0.1) (long dashed).

There were two major issues involved in this paper, and two conclusions made:

1. *Structure*, by which we mean the distribution of the cells. Our contribution to this issue is to model the geometric cell positions as the order statistics from Beta distributions. We found that there was structure in the distal colon cells, in that they were more concentrated at the bottom of the crypts. We showed that the results of the analyses were essentially the same whether the Beta distributions had common parameters or not.
2. *Coordinated response*, by which we mean a non-standard correlation of DNA damage between two regions of the colon as a function of geometric cell position. In our experiment, geometric cell position was available only for a small number of crypts. In our Bayesian framework, we generated the 'missing' cell positions simultaneously and directly from the Beta distribution, rather than the

more direct but less efficient one-cell-at-a-time strategy. This approach can be used more generally for the study of other cell position-based effects in colon carcinogensis in the presence of structure, as well as in other settings with this type of missing data. As part of the analysis, we showed, both theoretically and in our example, that it was at least possible for an analysis using nominal instead of geometric cell position to arrive at misleading conclusions. However, in the actual data, either means of measuring cell position led to similar results. Specifically, a negative relationship was observed between the DNA damage processes in the distal and proximal colon at all depths within the crypt for rats fed corn oil diets, while a positive relationship was observed near the base of the crypts for fish oil-fed rats. This may indicate that the diets, in determining the lumenal environment of the colon, affect the activation and delivery of the carcinogen to the colon cells. It appears the corn oil-rich diet may lead to a more localized application of the carcinogen, which could in principle intensify its effects. Further studies are needed to verify these hypotheses.

ACKNOWLEDGEMENTS

This work was supported in part by the Center for Environmental and Rural Health, Texas A&M University, and the following grants: NIH CA57030, CA74552, CA59034, CA61750, and P30-ES09106. We would like to thank the reviewers, whose useful comments have dramatically improved this paper.

APPENDIX A

Further details on metropolis–hastings algorithms

Since the complete conditionals cannot be found in closed form, Metropolis-Hastings steps are used to sample from the distribution of the geometric cell position hyperparameters in Sections 2 and 3 as well as the missing geometric cell positions themselves in Section 3. Some details of these steps are given here, first for sampling the geometric cell position hyperparameters, then for the missing geometric cell positions.

The candidate (a, b) at each iteration of the Gibbs sampler is obtained by sampling from independent Gamma $\{\alpha(a), \gamma(a)\}$ and Gamma$\{\alpha(b), \gamma(b)\}$ distributions, with $\alpha(a) = a^2/\sigma_{ab}^2$ and $\gamma(a) = a/\sigma_{ab}^2$. The parameters are chosen such that the mean of their distribution is the (a, b) from the previous iteration and the variance is σ_{ab}^2. The variance σ_{ab}^2 is chosen so that the probability of accepting a new candidate value is neither too small, nor too large. A candidate value that is too small results in many repeats in the posterior distribution, and one that is too large results in small moves that might fail to quickly explore the parameter space (Tanner, 1996).

The acceptance probability for the candidate values, $\theta_{\text{new}} = (a_{d,\text{new}}, b_{d,\text{new}})$ is

$$\alpha(\theta_{\text{new}}, \theta_{\text{old}}) = \min\left\{1, \frac{g(\theta_{\text{new}})q(\theta_{\text{new}}, \theta_{\text{old}})}{g(\theta_{\text{old}})q(\theta_{\text{old}}, \theta_{\text{new}})}\right\} \tag{A.1}$$

$$\text{where } g(a, b) = a^{\alpha_a - 1} b^{\alpha_b - 1} \exp(-\gamma_a a + \gamma_b b)\{\beta(a, b)\}^{-N} \prod_{ijrc_dl} (X_{ijrc_dl})^{a-1}(1 - X_{ijrc_dl})^{b-1},$$

with $N = \sum_{ijrc_d} n_{ijrc_d}$ and $q(\theta_1, \theta_2)$ the probability of moving from θ_1 to θ_2. The function q is the product of two densities, namely the density of a Gamma$\{\alpha(a_1), \gamma(a_1)\}$ evaluated at a_2 and the density of a Gamma$\{\alpha(b_1), \gamma(b_1)\}$ evaluated at b_2.

Using the Metropolis-Hastings algorithm, the candidate vector of geometric cell positions for each missing distal crypt c_d of rat r from time group j, fed a diet supplemented with corn oil is obtained by taking the order statistics of a random sample of size n_{2jrc_d} from a beta(a_d, b_d) distribution. Candidate

values for proximal crypts are found in an equivalent manner. The acceptance probability for the candidate vector is

$$\alpha(X_{\text{old}}, X_{\text{new}}) = \min\{1, g_X(X_{\text{new}})/g_X(X_{\text{old}})\}, \tag{A.2}$$

where $g_X(X) = \exp[-(2 * \sigma^2_{\epsilon_d})^{-1} \sum_l \{Y_{jrc_dl} - Z(X)\beta_{jrc_d}\}^2]$.

APPENDIX B

Complete conditional distributions for Bayesian analysis

The Bayesian methodology in Section 3 requires the calculation of the complete conditionals for all fixed and covariance parameters in the model, which can be done by standard calculations involving Gaussian and Wishart conjugate priors in the setting of the multivariate normal. Here we briefly present those complete conditionals.

With three rats per treatment group, the complete conditional of the vector of coefficients at the time level $(\beta'_{j,p}, \beta'_{j,d})'$ is multivariate normal with covariance $\Sigma^*_R = (3\Sigma_R^{-1} + \tau^{-1})^{-1}$ and mean vector $\Sigma^*_R\{\Sigma_R^{-1}(\beta'_{j\cdot,p}, \beta'_{j\cdot,d})' + \tau^{-1}(\mu'_p, \mu'_d)'\}$, where $\beta_{j\cdot,d} = \sum_{r=1}^3 \beta_{jr,d}$, assuming the Gaussian prior for $(\beta'_{j,p}, \beta'_{j,d})'$ with mean (μ_p, μ_d) and covariance matrix τ. Σ_R is the rat-level covariance matrix, as defined in Section 3.3.

Assume both the distal and proximal crypt level parameter vectors are of length p and recall that $C_{jr,p}$ and $C_{jr,d}$ denote respectively the numbers of proximal and distal crypts for rat r, in time group j, fed a corn-oil-supplemented diet. Also let $\beta_{jr\cdot,p} = \sum_{c_p} \beta_{jrc_p,p}$ and $\Sigma_C = \text{diag}(\Sigma_{C,p}, \Sigma_{C,d})$, a $2p \times 2p$ block diagonal matrix with $\Sigma_{C,p}$ and $\Sigma_{C,d}$ defined in Section 3.3. The coefficients at the rat level, $(\beta'_{jr,p}, \beta'_{jr,d})'$ also have a multivariate normal complete conditional, with covariance $\Sigma^*_C = [\{\text{diag}(C_{jr,p}, C_{jr,d}) \otimes I_p\}\Sigma_C^{-1} + \Sigma_R^{-1}]^{-1}$ and mean vector $\Sigma^*_C\{\Sigma_C^{-1}(\beta'_{jr\cdot,p}, \beta'_{jr\cdot,d})' + \Sigma_R^{-1}(\beta'_{j,p}, \beta'_{j,d})'\}$, where I_p is the $p \times p$ diagonal matrix and $\otimes$ denotes the Kronecker product.

The coefficients at the crypt level in the proximal region $\beta_{jrc_p,p}$ have a multivariate normal complete conditional with covariance $\Sigma^*_E = \{(\sum_l X'_{jrc_pl,p} X_{jrc_pl,p})/\sigma^2_{\epsilon,p} + \Sigma_C^{-1})^{-1}$ and mean vector $\Sigma^*_E\{\sigma^{-2}_{\epsilon,p}(\sum_l X'_{jrc_pl,p} Y_{jrc_pl,p}) + \Sigma_C^{-1}(\beta_{jr,p})\}$, with Σ_C defined as above and $\sigma^2_{\epsilon,p}$ defined in Section 3.3. Distal conditionals at the crypt level are defined similarly.

For $\sigma^2_{\epsilon,p}$, assuming an Inverse Gamma prior with parameters $\alpha_{\epsilon,p}$ and $\gamma_{\epsilon,p}$, the complete conditional distribution is Inverse Gamma with parameters $\alpha_{\epsilon,p} + N_p/2$ and $\{\gamma^{-1}_{\epsilon,p} + \sum_{jrc_pl}(Y_{jrc_pl,p} - X_{jrc_pl,p}\beta_{jrc_p,p})^2/2\}^{-1}$, assuming N_p is the total number of proximal observations used in the analysis. The distal results follow similarly.

The complete conditionals for the covariance matrix at the rat level Σ_R and the covariance matrices from the proximal and distal crypt levels $\Sigma_{C,p}$ and $\Sigma_{C,d}$, respectively, are all Inverse Wisharts, whose density is given in Muirhead (1982, page 97), assuming their priors are all Inverse Wisharts with parameters (α_R, Γ_R), $(\alpha_{C,p}, \Gamma_{C,p})$, and $(\alpha_{C,d}, \Gamma_{C,d})$, respectively.

For Σ_R, a $2p$ x $2p$ matrix, the complete conditional is Inverse Wishart of dimension $2p$ with parameters α^*_R and Γ^*_R, denoted $\text{Wi}^{-1}_{2p}(\alpha^*_R, \Gamma^*_R)$, with $\alpha^*_R = \alpha_R + R + 2p - 1$, where R=total number of rats in analysis, and $\Gamma^*_R = \{\Gamma_R + \sum_{jr}(\beta_{jr} - \beta_j)'(\beta_{jr} - \beta_j)\}^{-1}$, defining $\beta_{jr} = (\beta'_{jr,p}, \beta'_{jr,d})'$ and $\beta_j = (\beta'_{j,p}, \beta'_{j,d})'$. For $\Sigma_{C,p}$, a p x p matrix, the complete conditional is $\text{Wi}_p^{-1}(\alpha^*_{C,p}, \Gamma^*_{C,p})$, with $\alpha^*_{C,p} = \alpha_{C,p} + \sum_{jr} C_{jr,p} + p - 1$ and $\Gamma^*_{C,p} = \{\Gamma_{C,p} + \sum_{jrc_p}(\beta_{jrc_p,p} - \beta_{jr,p})'(\beta_{jrc_p,p} - \beta_{jr,p})\}^{-1}$, with the distal defined similarly.

REFERENCES

BARNES, C. J., HARDMAN, W. E. AND CAMERON, I. L. (1999). Presence of well-differentiated distal, but not poorly differentiated proximal, rat colon carcinomas is correlated with increased cell proliferation in and lengthening of colon crypts. *International Journal of Cancer* **80**, 68–71.

CARROLL, R. J., RUPPERT, D. AND STEFANSKI, L. A. (1995). *Measurement Error in Nonlinear Models*. London: Chapman and Hall.

CHANG, W. C. L., CHAPKIN, R. S. AND LUPTON, J. R. (1997). Predictive value of proliferation, differentiation and apoptosis as intermediate markers for colon tumorigenesis. *Carcinogenesis* **18**, 721–730.

FIALA, E. S. (1977). Investigations into the metabolism and mode of action of the colon carcinogens 1,2-Dimethylhydrazine and Azoxymethane. *Cancer* **40**, 2436–2445.

FULLER, W. A. (1987). *Measurement Error Models*. New York: Wiley.

GILKS, W. R., RICHARDSON, S. AND SPIEGELHALTER, D. J. (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

GELMAN, A. AND RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457–472.

HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

HONG, M. Y., LUPTON, J. R., MORRIS, J. S., WANG, N., CARROLL, R. J., DAVIDSON, L. A., ELDER, R. H. AND CHAPKIN, R. S. (2000). Dietary fish oil reduces O^6-methylguanine DNA adduct levels in the rat colon in part by increasing apoptosis during tumor initiation. *Cancer Epidemiology, Biomarkers and Prevention* **9**, 819–826.

LITTLE, R. J. A. AND RUBIN, D. B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. AND TELLER, E. (1953). Equations of state calculations by fast computating machines. *Journal of Chemical Physics* **21**, 1087–1092.

MORRIS, J. S., WANG, N., LUPTON, J. R., CHAPKIN, R. S., TURNER, N. D., HONG, M. Y. AND CARROLL, R. J. (2001). Parametric and nonparametric methods for understanding the relationship between carcinogen-induced DNA adduct levels in distal and proximal regions of the colon. *Journal of the American Statistical Association* **96**, 816–826.

MUIRHEAD, R. J. (1982). *Aspects of Multivariate Statistical Theory*. New York: Wiley.

PARZEN, E. (1960). *Modern Probability Theory and its Applications*. New York: Wiley.

RAMSAY, J. O. AND SILVERMAN, B. W. (1997). *Functional Data Analysis*. New York: Springer.

ROGERS, K. J. AND PEGG, A. E. (1977). Formulation of 06-Methylguanine by alkylation of rat liver, colon, and kidney DNA following administration of 1,2-Dimethylhydrazine. *Cancer Research* **37**, 4082–4087.

SATO, M. AND AHNEN, D. J. (1992). Regional variability of colonocyte growth and differentiation in the rat. *Anatomical Record* **233**, 409–414.

SHIH, I. M., WANG, T. L., TRAVERSO, G., ROMANS, K., HAMILTON, S. R., BEN–SASSON, S., KINZLER, K. W. AND VOGELSTEIN, B. (2001). Top–down morphogenesis of colorectal tumors. *Proceedings of the National Academy of Science* **98**, 2640–2645.

TANNER, M. A. (1996). *Tools for Statistical Inference: Models for the Exploration of Posterior Distributions and Likelihood Functions*, 3rd edn. New York: Springer.

WANG, N. AND DAVIDIAN, M. (1996). A note on covariate measurement error in nonlinear mixed effects models. *Biometrika* **83**, 801–812.

[*Received October 20, 2000; first revision May 2, 2001; second revision October 8, 2001; accepted for publication November 19, 2001*]