BIOINFORMATICS

Vol. 22 no. 14 2006, pages e332–e340
doi:10.1093/bioinformatics/btl239

# Mutation parameters from DNA sequence data using graph theoretic measures on lineage trees

Reuma Magori-Cohen[1], Yoram Louzoun[1] and Steven H. Kleinstein[2,*]

[1]Math Department, Bar Ilan University, Ramat Gan, Israel, 52900 and [2]Department of Computer Science, Princeton University, Princeton, New Jersey, 08544, USA

## ABSTRACT

**Motivation:** B cells responding to antigenic stimulation can fine-tune their binding properties through a process of affinity maturation composed of somatic hypermutation, affinity-selection and clonal expansion. The mutation rate of the B cell receptor DNA sequence, and the effect of these mutations on affinity and specificity, are of critical importance for understanding immune and autoimmune processes. Unbiased estimates of these properties are currently lacking due to the short time-scales involved and the small numbers of sequences available.

**Results:** We have developed a bioinformatic method based on a maximum likelihood analysis of phylogenetic lineage trees to estimate the parameters of a B cell clonal expansion model, which includes somatic hypermutation with the possibility of lethal mutations. Lineage trees are created from clonally related B cell receptor DNA sequences. Important links between tree shapes and underlying model parameters are identified using mutual information. Parameters are estimated using a likelihood function based on the joint distribution of several tree shapes, without requiring *a priori* knowledge of the number of generations in the clone (which is not available for rapidly dividing populations *in vivo*). A systematic validation on synthetic trees produced by a mutating birth-death process simulation shows that our estimates are precise and robust to several underlying assumptions. These methods are applied to experimental data from autoimmune mice to demonstrate the existence of hypermutating B cells in an unexpected location in the spleen.

**Contact:** stevenk@cs.princeton.edu

## 1 INTRODUCTION

Mutating birth-death processes (MBDPs) are a fundamental component of biology at many different time scales, ranging from evolution of species, through epidemiological evolution of bacteria and other pathogens, to within-host mutation of viruses such as HIV. A special case is the affinity maturation of B cells during an immune response, which is the main focus of this paper (although our methods could be applied to other MBDPs). Affinity maturation normally occurs following the migration of naïve B cells into germinal centers, and the binding of B cell antibody receptors directly to antigens (e.g., molecular determinants on the surface of pathogens) accompanied by a secondary signal resulting from binding helper T cells. Over a three-week period, activated B cells proliferate rapidly and undergo a process of somatic hypermutation whereby point mutations are introduced into the DNA coding for their antibody receptor. According to the theory of clonal selection, B cells with mutations that increase their affinity for antigen gain a proliferative advantage. In this way the average affinity of the population increases over time. For a detailed review of the biology underlying affinity maturation, please see (Wagner and Neuberger 1996).

Despite the significance of MBDPs, methods to estimate underlying parameters from available data are lacking in many cases, particularly when the number of samples is small and the time-scale is short (as is the case for B cell affinity maturation). Population genetic methods have been developed to estimate various evolutionary parameters based on constant rate birth-death models (Nee *et al*., 1994) or coalescent processes (Rosenberg and Nordborg 2002). These approaches assume a large population size (often fixed in the case of coalescent processes) and superimpose a mutation history on a genealogy as a separate step. However, when the number of generations is small the mutation rate impacts the tree topology along with the branch lengths. In addition, population genetic models do not include processes, such as mutation-dependent cell death, that play an important role in B cell affinity maturation.

We previously reported preliminary results on maximum likelihood (ML) methods to estimate the B cell receptor mutation rate based on a small number of B cell lineage tree shapes (Kleinstein *et al*., 2003). Each tree is obtained from a microdissection experiment, which provides a number of clonally related B cell receptor DNA sequences that can be genealogically related to each other using a maximum parsimony algorithm (Clement *et al*., 2000). Many processes, including the hypermutation rate, influence the 'shape' of these clonal trees. Our approach defines a simple MBDP comprising the main biological mechanisms underlying affinity maturation (including clonal expansion with somatic hypermutation and the possibility of lethal mutations), and estimates the parameters of this process. Applying this to biological data from at the Shlomchik lab (Kleinstein *et al*., 2003) suggested, for example, that specific B cells in an autoimmune mouse were undergoing somatic hypermutation in an unexpected area of the spleen. Here we extend and confirm these findings, which have important implications for understanding the etiology of autoimmune diseases such as Lupus.

*To whom correspondence should be addressed.

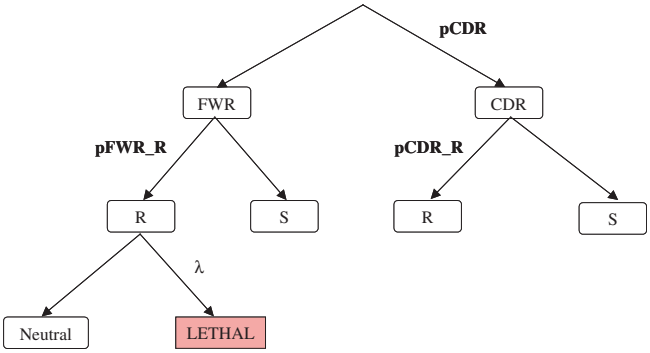


**Fig. 1.** Mutation decision tree used in MBDP simulations. Mutations occur with a Poisson rate of $\mu$ per division. The effect of each mutation depends on whether it falls in the Framework Region (FWR) or the Complementarity Determining Region (CDR) of the receptor gene. The relatively invariant FWRs provide the overall structure of the receptor, and serve to support the more variable CDRs where antigen binding commonly occurs. Following the work of (Shlomchik, Watts *et al*. 1998), each mutation was given a pCDR = 25% chance of being a CDR mutation and a (1-pCDR) = 75% chance of being a FWR mutation. Each mutation also has a pCDR_R = pFWR_R = 75% chance of being a replacement and (1-pCDR_R) = (1-pFWR_R) = 25% chance of being silent. FWR replacement mutations have a probability $\lambda$ of being lethal. Lethal mutations kill the cell.

The parameters of the MBDP would ideally be estimated by computing the likelihood of producing the set of observed lineage trees (or isomorphic ones) over all realistic parameter values, and finding the combination producing the highest overall probability. However, such full ML methods are expensive to compute. Our previous work sought to develop a more computationally efficient method by summarizing each tree as a set of graph theoretic measures (referred to as tree shapes). In this study we use mutual information to focus the analysis on the most informative shapes. Previous attempts to study tree shapes in the context of evolutionary processes have usually focused on single properties such as the distribution of the number of lineages over time (Nee *et al*., 1994), or branch lengths (Takezaki *et al*., 1995). Analysis of B cell lineage trees up to this point has been limited to statistical approaches based on the average values of individual tree shapes, and qualitative comparison of possible underlying processes (Dunn-Walters *et al*., 2002; Dunn-Walters *et al*., 2004; Mehr *et al*., 2004). The current work differs from these previous attempts in the development of an underlying stochastic model appropriate for B cell clonal expansion, and the quantitative correlation of several tree shapes to allow estimation of multiple parameters of the MBDP model.

## 2 MODEL AND METHODS

B cell clonal expansion and somatic hypermutation are modeled as a MBDP with multi-type cells and the following three reactions:

- Cell division with average rate of $\beta$ per generation.
- Stochastic mutation with Poisson rate $\mu$ per division.
- Cell death with probability of $\lambda^*$ per mutation, where $\lambda^* = (1.0 - \text{pCDR}) * \text{pFWR\_R} * \lambda$. pCDR, pFWR and $\lambda$ are parameters of the mutation decision tree (Figure 1).
- Mutation-independent cell death with rate $\delta$ per division.

**Table 1.** Tree shapes considered in the mutual information analysis. Shapes used in the simulation-based estimate (S) and the analytical estimate (A) are indicated. 'Full' vertices contain observed sequences (seq), while 'empty' ones do not

| | Tree Shape Description | S | A |
|---|---|---|---|
| 0 | Number of full internal vertices | X | |
| 1 | Number of empty internal vertices | X | |
| 2 | Sequence in full internal vertices | | |
| 3 | Number of parent-child couples | | |
| 4 | # of seq. in parent-child couples | | X |
| 5 | # of seq. in vertices with empty parent | | |
| 6 | Number of repeated sequences | | X |
| 7 | Number of internal vertices | X | |
| 8 | Number of leaves | | |
| 9 | Number of seq. in leaves | | |
| 10 | Number of seq. in internal vertices | | |
| 11 | Number of vertices | X | |
| 12 | Number of sequences at root* | X | X |
| 13 | Number of edges | | |
| 14 | Number of independent mutations | | |
| 15 | Average number of mutations* | X | X |
| 16 | Least common ancestor distance | | |
| 17 | Replacement-to-Silent ratio, R/(R+S) | X | X |

*Note that the simulation estimate defines these measures on unique sequences only.

The MBDP is initiated with a single cell. After $d$ cell generation times, $q$ cells are randomly sampled from the total population of $N$ cells. The set of accumulated mutations in the $q$ sampled cells are used to create a genetic lineage tree as previously described (Kleinstein *et al*., 2003). By construction this tree is correct (i.e., it is a sub-tree of the actual lineage tree). We assume that the experimentally observed trees (created using maximum parsimony on sets of B cell receptor DNA sequences) are correct in the same sense.

The parameters of the MBDP are $\theta = (\beta,\mu,\lambda,\delta,\text{d},q)$. The division rate ($\beta$), assumed to be equal for all cells, defines the time scale of the MBDP and can be set to one through appropriate rescaling. Throughout this study we set $\beta = 1$ and simulate cell division as a deterministic process occurring once during each discrete time step of the simulation. As shown later, relaxing this deterministic assumption has only a minor impact on our estimates. Our estimation methods also assume $\delta = 0$, although we include this parameter in the analytical formula derivations and mutual information analysis. We don't expect this to affect the other parameter estimates (see Summary). The number of sampled cells ($q$) is included as a parameter to account for the possibility that some observed sequences are repetitions from a single sampled cell due to the particular experimental protocols used. While the simulation-based method we present is insensitive to these potential repetitions, the analytical method assumes a one-to-one correspondence between sampled cells and observed sequences.

Each lineage tree $t$ is summarized by a set of shapes: $S_t = \{s_{t,1}, s_{t,2}, \ldots s_{t,S}\}$, where $s_{t,i}$ is shape i of tree t. The shapes used in this study (defined in Table 1) partially overlap those defined in (Kleinstein *et al*., 2003). We estimate $\theta$ by maximizing the likelihood of observing the set of tree shapes $S_t$ over all trees t given the MBDP described above with parameters $\theta$. Expected tree shapes are based on analytical formulae or numerical simulations. To maximize the available information, our approach uses the collective properties of a set of trees assuming the same generative process and equivalent $\mu$ and $\lambda$, but different d and $q$ for each tree.

We use synthetic data sets to measure the precision of these estimates under different experimental conditions and show that our methods work

with realistic (i.e., small) amounts of experimental data. We first present and validate the methods on synthetic data and then apply them to a set of B cell receptor sequence data derived from microdissection experiments in a mouse model of autoimmune disease. While results from our previous analysis were limited to the mutation rate, here we extend and validate the methods to estimate additional parameters, specifically the lethal mutation frequency and the number of divisions in each clone.

## 2.1 Simulation of B-cell clonal expansion

The MBDP simulation has been previously described (Kleinstein *et al*., 2003; Kleinstein and Singh 2003). Briefly, it is initiated with a single seeding cell. At each discrete time step (corresponding to one cell generation time), all cells are allowed to divide and die. During each division a Poisson distributed number of mutations occurs with average $\mu$. The impact of a mutation is determined by the mutation decision tree in Figure 1. Cells with lethal mutations are removed after every generation. This process continues until $D_{max}$ generations have passed.

## 2.2 Mutual information analysis

Mutual information based on Shannon's entropy, one of the central concepts of Information theory, is used to identify tree shapes (both individually and in groups) that provide the maximal information about the underlying MBDP parameters. Shannon's entropy measures the information content of a source X, and is defined as:

$$H(X) = -\sum_{x} \Pr(x) \cdot \log \Pr(x)$$

where Pr(x) is the probability function of the random variable X. Similarly, joint Shannon's entropy is defined as:

$$H(X,Y) = -\sum_{x,y} \Pr(x,y) \log \Pr(x,y)$$

where Pr(x,y) is the joint probability function of the random variables X and Y. These formulas are used to calculate the information content of each combination of tree shapes, denoted as H(Y), as well as combinations of MBDP parameters, denoted as H(X). By definition, the mutual information between the parameters and shapes is:

$$I(X,Y) = H(X) + H(Y) - H(X,Y)$$

The mutual information measures the information about *X* that is shared by *Y*. In other words, how much of the information about the model parameters X is expressed by the tree shapes Y.

## 2.3 Analytical estimation method

In the analytical method, optimal model parameters are estimated by minimizing the weighted least squares difference between the observed and expected tree shapes:

$$X(\theta) = \sum_{t=1}^{T} \underset{d}{Min} \left( \sum_{i=1}^{S} \frac{(s^{o}_{t,i} - s^{e}_{t,i}(\theta))^2}{VAR(\{s^{o}_{r,i}\}_{r=1}^{T})} \right)$$

where S is the number of tree shapes considered, $s^{o}_{t,i}$ is the observed value of shape i in tree t, and $s^{e}_{t,i}(\theta)$ is the expected value given the parameters $\theta$. $VAR(\{s^{o}_{r,i}\}_{r=1}^{T})$ is the variance of tree shape *i* calculated over all the observed lineage trees. The minimization of the error $X(\theta)$ takes place in two stages. For each observed tree t, we first minimize the error over all possible numbers of divisions (d), producing an estimate for the number of divisions in the clone that gave rise to the tree (denoted $d_t$). The overall error is then computed as the sum of the errors for each tree, and this value is minimized to find the optimal values for $\mu$ and $\lambda$ simultaneously. Recall that in this approach we assume the number of sampled cells ($q$) equals the observed number of sequences as discussed above.

## 2.4 Simulation-based estimation method

In the simulation-based approach, we begin by estimating $\lambda$ to be the value where the expected fraction of mutations that are replacements, R/(R+S), is equal to its observed value ($I_t$) computed over all independent mutations in all trees:

$$\lambda = (I_t - (FWR\_R + CDR\_R))/(FWR\_R*(I_t - 1))$$

where,

$$FWR\_R = (1.0 - pCDR)*pFWR\_R$$
$$CDR\_R = pCDR*pCDR\_R$$

pCDR, pCDR_R and pFWR_R are parameters of the mutation decision tree shown in Figure 1, which describes the distribution of random mutations. These parameters are set to typical values (Shlomchik, Watts *et al*., 1998) although, as we have previously shown, it is easy to estimate them directly for any particular germline sequence of interest (Kleinstein and Singh 2003).

The overall likelihood for producing an experimental data set is the product of the likelihood for each observed tree:

$$L(S_1, S_2 \ldots S_T \mid \theta) = \prod_{t=1}^{T} L(S_t \mid u_t, \theta)$$

where $L(S_t \mid u_t,\theta)$ is the likelihood of observing a tree with shapes $S_t$ given that the microdissection and sequencing produces $u_t$ unique sequences, and assuming an underlying model with parameters $\theta$. This likelihood is also dependent on the number of divisions in the clone, and the number of cells sampled to create the tree. Since neither of these quantities are known with certainty, we sum over all possible values (assuming they are equally likely) to get:

$$L(S_t \mid u_t, \theta) = \sum_{d=1}^{\infty} \sum_{q=u_t}^{q_t} \Pr(S_t, d, q \mid u_t, \theta)$$

where we know that the number of sampled cells included in tree t lies somewhere between the observed number of unique sequences ($u_t$) and the total number of cells in the microdissection pick ($q_t$). Of course we cannot simulate an infinite number of divisions in practice, and it is necessary to limit the number of divisions to a computationally reasonable range. This can lead to errors in the estimate due to truncation of the distribution so instead of summing over the entire range we choose the value of d for each tree that maximizes the probability (referred to as $d_t$):

$$L(S_t \mid u_t, \theta) = \underset{d \le D_{max}}{Max} \sum_{q=u_t}^{q_t} \Pr(S_t, d, q \mid u_t, \theta)$$

The maximum ($d_t$) is our estimate for the number of generations in the clone that gave rise to tree t. The upper bound on the number of divisions in the simulation ($D_{max}$) is set to a value that is thought to upper bound the clonal expansion size (and is computationally feasible). $\Pr(S_t,d,q \mid u_t,\theta)$ is estimated using Monte Carlo simulations with parameters $\theta$ by:

$$\Pr(S_t, d, q \mid u_t, \theta) = \frac{E(S_t, d, q, u_t)}{\sum_{d,q} U(d, q, u_t)}$$

Here, $U(d,q,u_t)$ is the number of simulated trees with $u_t$ unique sequences after *d* divisions and randomly sampling *q* cells. Among these trees, $E(S_t,d,q,u_t)$ is the number that are also equivalent to the observed tree in all shapes specified in Table 1 (i.e., the simulated tree can be summarized by $S_t$).

To calculate $U(d,q,u_t)$ and $E(S_t,d,q,u_t)$, an expanding B cell clone is simulated beginning from a single cell as described in Section 2.1. After each division, *q* cells are randomly sampled from the population and a lineage tree is created as previously described (Kleinstein *et al*., 2003). If the number of unique sequences in this tree is $u_t$, then $U(d,q,u_t)$ is incremented by one, where d is the number of divisions so far. If the simulated tree also has shape $S_t$, then $E(S_t,d,q,u_t)$ is incremented by one.

After running the simulation *i* times, the likelihood of producing each of the observed trees is determined. We used $i = 128{,}000$ independent simulation runs to calculate the likelihood at each value of the mutation rate since more runs did not provide additional accuracy (data not shown). The overall mutation rate $\mu$ is estimated by maximizing the likelihood using golden section search.

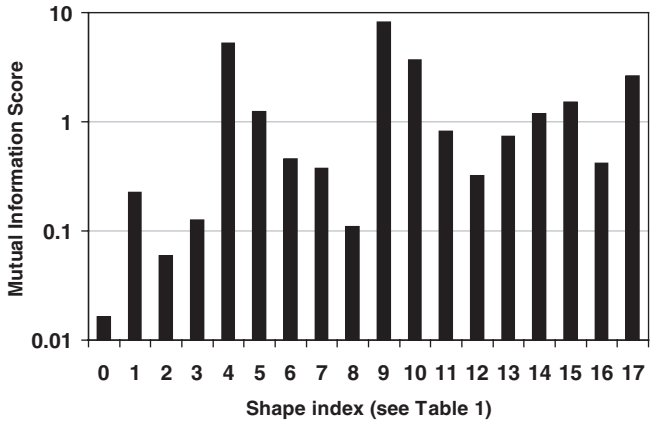


**Fig. 2.** Contribution of individual tree shapes to estimation of simulation parameters $\theta$. Shape triplets were ordered by their mutual information with the MBDP parameters, and weighted by an exponentially decreasing function. Bar heights are the sums of the weighted frequency of each shape. Synthetic data included 1000 trees for every combination of: $\mu = \{0.1, 0.3, 0.5, 0.7, 0.9\}, \lambda = \{0.0, 0.25, 0.50\}, \delta = \{0.0, 0.2, 0.4\}, \mathrm{d} = \{0,1,\ldots,15\}$, with $q = \min(\mathrm{N},10)$.

## 3 RESULTS

### 3.1 Optimal tree shapes

Individual tree shapes can reflect different aspects of the underlying biological process, but some contain similar information. The relationship between model parameters ($\theta$) and resulting tree shapes can be highly non-linear so that classical linear regression measures do not properly represent the contribution of particular shapes to the estimate of $\theta$. We use mutual information based on Shannon's entropy to determine which shapes contain the most information about $\theta$. By varying the set of shapes included in $S_t$ and measuring the mutual information $I(S_t,\theta)$, we can estimate how much information is conveyed by $S_t$ about $\theta$. We define the most informative set of shapes as the one with the highest mutual information.

To determine the most informative set of tree shapes, we computed the mutual information between $S_t$ and $\theta = (\mu,\lambda,\delta,\mathrm{d},q)$ for a set of synthetic trees produced by simulating the MBDP for a range of realistic parameter values (see Figure 2 caption). An equivalent analysis was done under the assumption that $q$ is known, using an extended shape vector $S_t^* = \{S_t,q_t\}$ and removing $q$ from $\theta$, and similar results were obtained (data not shown). We considered all possible combinations of three tree shapes from the set listed in Table 1. The optimal shape triplet for the simultaneous estimate of $\theta$ is composed of: (1) the ratio of Replacement (R) to Silent (S) mutations, (2) the number of sequences in parent-child couples, and (3) the number of sequences in the leaves of the tree. However, several other triplets have similar mutual information. We developed a scoring system to identify individual tree shapes that repeatedly appear in high mutual information triplets. First, all shape triplets are sorted by their mutual information content and given an exponentially decreasing weight. The score for individual shapes (i.e., components of $S_t$) is calculated by summing the weights of all triplets that contain it. As shown in Figure 2, the highest scoring individual shapes include those in the optimal triplet. Note that some shapes score highly, but are equivalent to other shapes so that there is no benefit to using them simultaneously, while other measures may have a relatively low score but are required to complete a good triplet. Quadruplets can be done in a similar way.

As described in the following section, we have derived analytical equations to approximate the expected values of several high scoring tree shapes in Figure 2 (indicated in Table 1). These shapes are used to summarize each lineage tree in the analytical estimation approach. The simulation-based estimate does not use many of these shapes since they would require assuming a one-to-one correspondence between sampled cells and observed sequences. We can avoid this assumption by restricting the set of valid shapes to those that do not depend on the number of repeated sequences. The limited number of such topological tree shapes allows us to use them all (up to closely related ones).

### 3.2 Analytical method results

As indicated in Table 1, five tree shapes were used in the error function to estimate $\theta$. The following sections outline formulas for the expected values for each of these tree shapes. The full development cannot be included here due to space limitations. Note that these derivations include mutation-independent cell death with rate $\delta$ per division.

*Average mutations per sequence.* Consider a cell that has undergone $d$ divisions and accumulated $m$ mutations. The number of such cells surviving (e.g., accumulating no lethal mutations) is: $\alpha^{\mathrm{d}}(1\text{-}\lambda_1)^{\mathrm{m}}$ where $\lambda_1$ = (1-pCDR) x pFWR_R x $\lambda$ is the overall probability that a mutation will be lethal (see Figure 1), and $\alpha = 2e^{-\delta}$. The probability of having $m$ mutations after $d$ generations is a Poisson process with an average of $\mu$d. Thus, after $d$ divisions the expected number of cells with $m$ mutations that are still alive is:

$$\alpha^d(1-\lambda_1)^{\mathrm{m}}e^{-\mu d}\frac{(\mu d)^m}{m!}$$

The number of live cells (N) after $d$ divisions is calculated by summing over all possible numbers of mutations:

$$N = \sum_m \alpha^d(1-\lambda_1)^{\mathrm{m}}e^{-\mu d}\frac{(\mu d)^m}{m!}$$

Thus, the expected number of mutations per sequence is:

$$\begin{aligned}\mathrm{M} &= \frac{1}{N}\sum_m \alpha^d(1-\lambda_1)^{\mathrm{m}}e^{-\mu d}\frac{(\mu d)^m}{m!}m \\ &= \mathrm{E}[\text{Poisson process with mean}(1-\lambda_1)\mu d] \\ &= (1-\lambda_1)\mu d\end{aligned}$$

After simplifying, we find that M is simply the expected branch length in the absence of lethal mutations ($\mu$d) multiplied by the probability of cell survival ($1-\lambda_1$).

*Number of unique sequences.* The expected number of unique sequences (u) in a random sample of $q$ cells can be computed by the probability that a given sequence is different from all others. The number of unique sequences is:

$$\sum_{i=1}^{q}\Pr(\textit{sequence } \mathrm{i} \neq \textit{sequence } j; j < i)$$

This can be approximated by:

$$\sum_{i=1}^{q}\Pr(i \neq 1)\Pr(i \neq 2 \,|\, i \neq 1)\ldots\Pr(i \neq i-1 \,|\, i \neq 1..i-2)$$

The probability that two random sequences are different is:

$$1 - X, \text{ where } X = \frac{(e^{-2\mu}((2\cdot\alpha\cdot e^{-2\mu})^d - 1))/(2\cdot\alpha\cdot e^{-2\mu} - 1)}{(2\cdot\alpha^d - 1)/(2\cdot\alpha - 1)}$$

The contribution of the conditional probability varies between 1 and $2^{1-i}$, so that the number of unique sequences is bounded between:

$$u = \sum_{i=1}^{q}\prod_{j=1}^{i-1}(1-X)^i \text{ and } u = \sum_{i=1}^{q}\prod_{j=1}^{i-1}\left(1-\frac{X}{2^j}\right)^i.$$

Averaging these values gives an excellent fit to the number of unique sequences in simulated trees (data not shown).

*Average sequences at the root.* Each sequence appearing at the root of a lineage tree represents a cell that has undergone *d* divisions without accumulating any mutations. The probability of this occurring for a single cell is $e^{-\mu d}$. Such cells will be enriched in the population due to the death of cells that accumulate lethal mutations. The fraction of cells in the root is thus the expected number of unmutated cells divided by the total number of surviving cells:

$$\frac{\alpha^d e^{-\mu d}}{\alpha^d(1-\lambda_1\mu)^d} = \left(\frac{e^{-\mu}}{(1-\lambda_1\mu)}\right)^d$$

Multiplying this fraction by the number of sampled cells ($q_t$) in any particular clonal tree (t) gives the number of sequences expected to be present at the root ($R_t$):

$$R_t = q_t \times \left(\frac{e^{-\mu}}{(1-\lambda_1\mu)}\right)^d$$

*Sequences in parent-child nodes.* The probability for a pair of sequences to appear as a parent-child couple in a tree can be computed as the probability that in two nearby branches of the actual lineage tree, one branch is mutated while the other is not. When the tree is collapsed to create the equivalent of the maximum parsimony tree, the sequence in the unmutated branch becomes the parent of the sequence in the mutated branch. The probability to find two such sequences is:

$$\begin{aligned}
&\sum_{i=1}^{d} e^{-\mu i}(1-e^{-\mu i})\cdot(2\cdot e^{-\lambda\mu-\delta})^{2i-2}(2\cdot e^{-\lambda\mu-\delta})^{d-i}\cdot\frac{q(q-1)}{N(N-1)}\\
&\quad\times 2^{(1-\mu-5\delta)} = \frac{q(q-1)\cdot(2\cdot e^{-\lambda\mu-\delta})^{d-2}}{N(N-1)}\cdot 2^{(1-\mu-5\delta)}\cdot\\
&\quad\times\left[\sum_{i=1}^{d}(2e^{-\mu-\lambda\mu-\delta})^i - \sum_{i=1}^{d}(2e^{-2\mu-\lambda\mu-\delta})^i\right]\\
&\quad= \frac{q(q-1)\cdot(2\cdot e^{-\lambda\mu-\delta})^{-2}}{(N-1)}\cdot 2^{(1-\mu-5\delta)}\cdot\\
&\quad\times\left[\frac{2e^{-\mu-\lambda\mu-\delta}(2^d e^{-\mu d-\lambda\mu d-\delta d}-1)}{2e^{-\mu-\lambda\mu-\delta}-1}\right.\\
&\quad\left.-\frac{2e^{-2\mu-\lambda\mu-\delta}(2^d e^{-2\mu d-\lambda\mu d-\delta d}-1)}{2e^{-2\mu-\lambda\mu-\delta}-1}\right]
\end{aligned}$$

## Estimates using the analytical method

The direct application of the analytical error minimization method, using the tree shapes in Table 1 with expected value computations described above, provides unbiased estimates of $\mu$ and $\lambda$ as tested on synthetic data sets (Figure 3). Looking at the error surface, we initially found that the minimum was indeed at the correct location, but the error function was practically flat in the direction of $\lambda$, resulting in a large variance in the estimate of the frequency of lethal mutations. This was improved by increasing the weight of R/(R+S) in the error function (Figures 3 and 4). This tree shapes is directly affected by $\lambda$, but not by $\mu$.

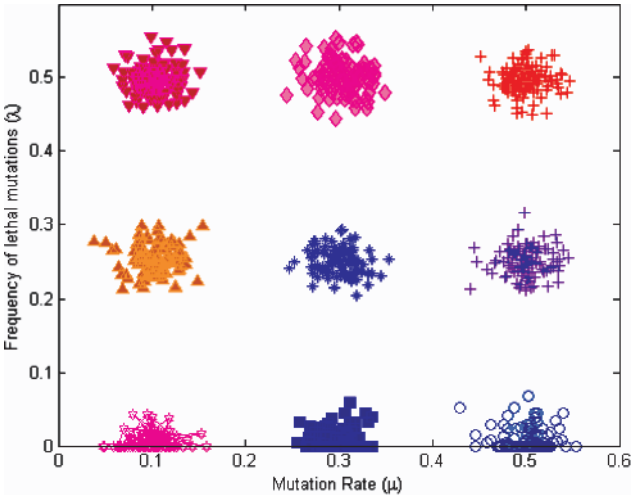


**Fig. 3.** Estimate of the mutation rate ($\mu$) and lethal frequency ($\lambda$) using the analytical method. Each synthetic data set contains T = 100 trees with $q = 10$ cells. Data was created for all combinations of $\mu = (0.1,0.3,0.5)$ and $\lambda = (0.00,0.25,0.50)$. Each point is the estimated value on one synthetic data set. Clusters are centered on the actual values. Note the Replacement-to-Silent ratio was given a 10-fold weight increase in the error function.

## 3.3 Simulation-based method results

As discussed previously (and shown in Figure 4 for the analytical method), the likelihood is relatively flat as a function of $\lambda$ suggesting that individual trees contain little information to estimate this parameter. Although we found it was possible to simultaneously estimate $\mu$ and $\lambda$ using the analytical approach, the inherent noise in the simulation estimate, resulting from the limited number of simulations used to estimate each likelihood, makes that strategy infeasible here. Thus, we employ a two-step approach. First the lethal mutation frequency ($\lambda$) is estimated by considering the fraction all independent mutations that are replacements, R/(R+S). Since silent mutations, which do not change the amino acid coded for, cannot be lethal to the cell, this fraction provides a direct, albeit noisy, signal to estimate the lethal mutation frequency. The mutation rate $\mu$ is then estimated by comparing the expected and observed tree shapes (not including R/(R+S) which was used to estimate $\lambda$), assuming a particular value of $\lambda$ (either known or estimated). Previous estimates of $\mu$ were based on educated guesses about the number of generations. Our method simultaneously estimates $\mu$ along with the number of divisions giving rise to each tree (further denoted $d_t$).

In contrast to the analytical method, the simulation-based method is robust to the assumption that repeated sequences represent unique cells. First, this method uses only shapes that do not depend on the precise number of repeated sequences. Second, all possible values for the number of sequences (q) are considered in the likelihood

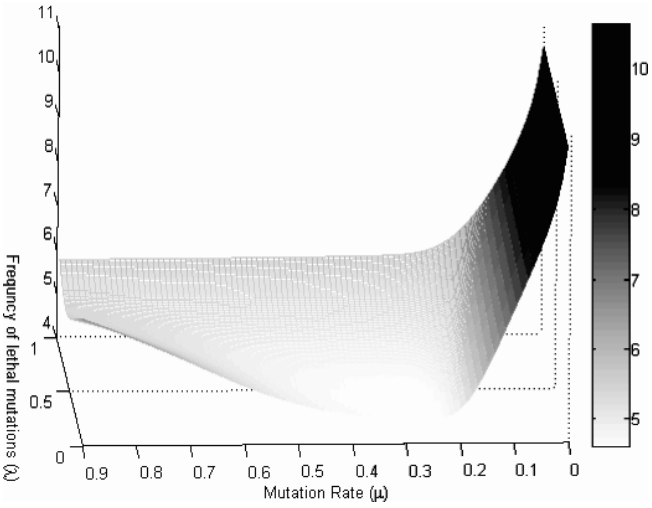


**Fig. 4.** Representative error landscape for analytical estimates based on a synthetic data set with 100 trees for $\mu$ and $\lambda$ where the actual values are $\mu = 0.3$ and $\lambda = 0.25$. The minimum is at the correct position, but the error landscape is very flat in the $\lambda$ direction. Note the Replacement-to-Silent ratio was given a 10-fold weight increase in the error function.

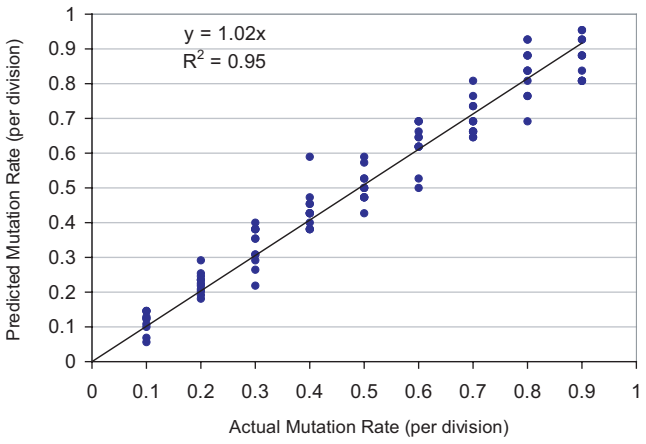


**Fig. 5.** Estimating the mutation rate with the simulation method when the frequency of lethal mutations ($\lambda$) is known. Individual points show the results on each synthetic data set, consisting of $T = 50$ clonal trees each with $q = 5$ cells. For each tree the sampling time was randomly distributed between 5 and 10 generations (as is the case with all our synthetic data sets except where indicated). At least 7 synthetic data sets were produced at each mutation rate investigated. $\lambda = 0.5$ and $D_{max} = 10$ for all likelihood evaluations.

computation. Another advantage of this approach is that, while the analytical formulas estimate the expected value of each tree shape independently, the simulation-based method numerically estimates the joint distribution of all shapes.

## The mutation rate and lethal frequency

In previous work we proposed a simulation-based method to estimate the mutation rate ($\mu$) from a set of lineage trees (Kleinstein *et al*., 2003). However, this method produced a biased estimate. The method presented here differs from that approach by explicitly summing over all possible numbers of sequences in the observed tree ($q_t$). Furthermore, instead of directly sampling the observed number of unique sequences from the simulated tree, we condition our likelihood on this value. This ensures that singletons (i.e., trees containing only a single unique sequence), whose true frequency is impossible to estimate experimentally, do not unduly influence the results.

We validate the improved method by estimating $\mu$ from synthetic data sets where the actual mutation rate is known. The frequency of lethal mutations ($\lambda$) has been previously estimated for some well-studied responses (Shlomchik 1990), and as a first step we consider the case where this parameter is known. From Figure 5, it is clear that this new method is unbiased and converges to the actual value of $\mu$. In addition, the variance in the estimate of the likelihood decrease as the sample size grows (data not shown). Even when the number of trees and sequences is small (as is the case for the actual experimental data), our method provides a reasonable estimate of the mutation rate (Figure 6).

Like the simulation used to estimate the likelihood function, the synthetic data sets assume that cell division is synchronous so that all cells in a clone (giving rise to a single tree) have undergone the same number of divisions. To test whether our method is sensitive to this assumption, we generated synthetic data sets with asynchronous division using the discrete time-step approach developed in

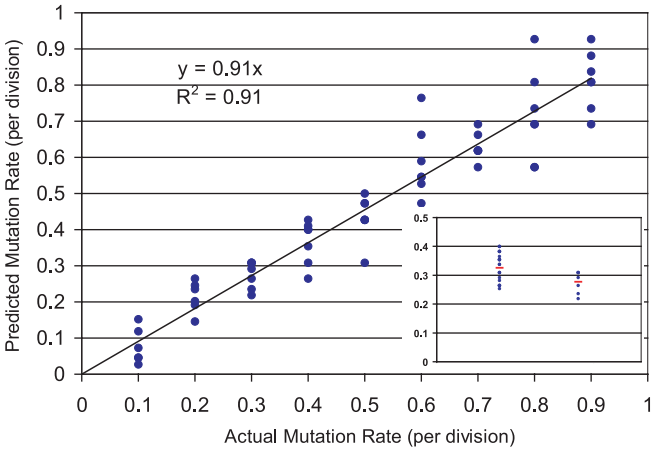


**Fig. 6.** Estimating the mutation rate for limited data with the simulation method when the frequency of lethal mutations ($\lambda$) is known. Synthetic data sets were created that had the same number of trees and sequences as the experimentally derived autoimmune response data. The inset shows that our method is not overly sensitive to the assumption of synchronous division. Synthetic data sets were created using an asynchronous division model (left) and a synchronous model (right). The simulation method (which uses a synchronous model) was applied to estimate the mutation rate whose actual value is $\mu = 0.3$ per division. $D_{max} = 12$ for all likelihood evaluations.

(Kleinstein and Singh 2001). In this model the time between divisions is Poisson distributed with average value $\beta = \ln(2)$ leading to a doubling time of one, which is equivalent to the synchronous model with discrete time steps. We then applied our estimation method (which still used synchronous division) to these data. The estimated mutation rates were close whether or not the synthetic data used synchronous or asynchronous division (Figure 6 inset).

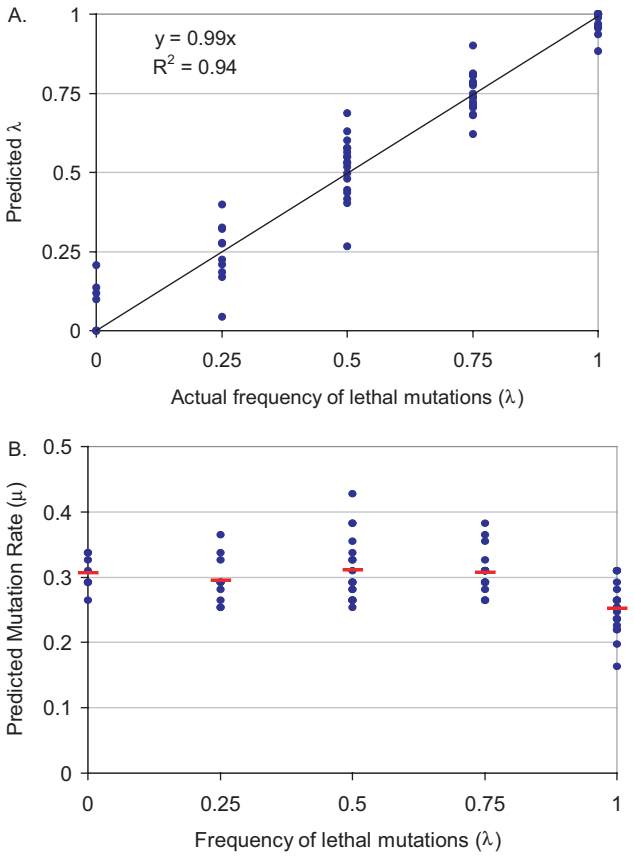


**Fig. 7.** Predicting the mutation rate with the simulation method when the frequency of lethal mutations ($\lambda$) is unknown. Synthetic data sets contain $T = 50$ trees with $q = 5$ cells each and $\mu = 0.3$. (A) The predicted value of lambda for individual synthetic data sets. (B) Predictions of the mutation rate for synthetic data sets created with the indicated value of $\lambda$. The underestimation at $\lambda = 1$ (which is biologically unrealistic in any case) is due to this being a hard upper bound. $D_{max} = 12$ for all likelihood evaluations.

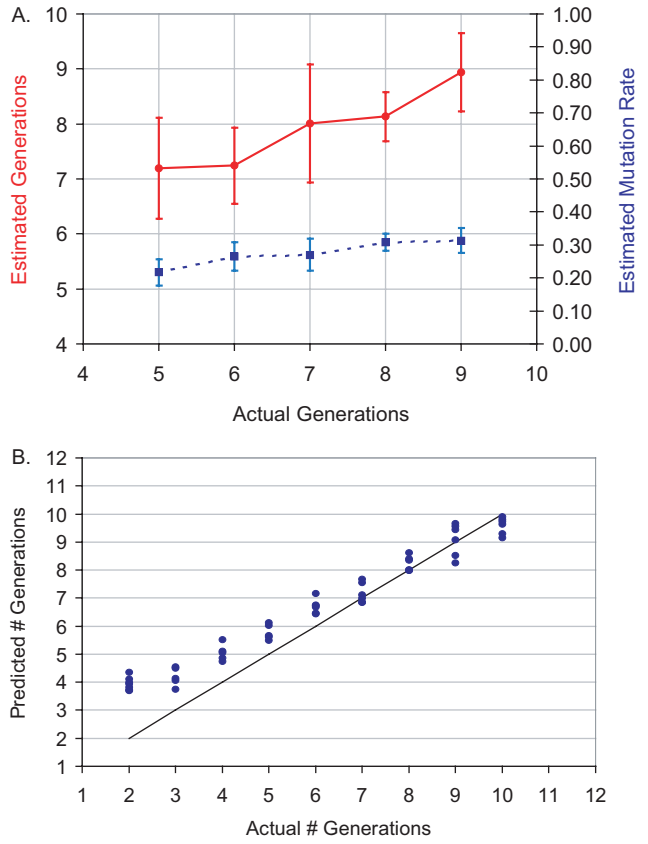


**Fig. 8.** Predicting the average number of generations using the simulation-based method. (A) Individual points are average results on at least 7 synthetic data sets consisting of $T = 50$ clonal trees each with $q = 5$ cells randomly sampled from a clone with the indicated number of generations. Error bars indicate one standard deviation. These results used the actual value for $\lambda$, and estimated the mutation rate $\mu$ (dotted line) and the average number of divisions for all clones in the data set (solid line). (B) The predicted number of generations when the mutation rate is known. Each point is the result from one synthetic data set. When the actual number of generations approaches the number of simulated generations used to calculate the likelihood ($D_{max}$), we can also underestimate the number of divisions. This is easily corrected by raising this bound at the expense of computation time. $D_{max} = 12$ for all likelihood evaluations.

We next consider how well we can estimate $\theta$ when the frequency of lethal mutations ($\lambda$) is unknown. Although we find that the estimate of $\lambda$ itself is quite noisy (Figure 7a), this does not greatly impact the estimate of $\mu$ (Figure 7b). In fact, when $\lambda = 0.5$ (a value estimated in previous work), the standard deviation in the prediction of $\mu$ increases only from 0.06 to 0.07 per division for the estimates with known and unknown $\lambda$ respectively. This is true even though the estimate of $\lambda$ itself has a standard deviation of 0.1 for the reasons previously discussed for the analytical estimate.

## The number of generations

To check if the number of divisions maximizing the likelihood for each tree ($d_t$) is a reasonable estimate of the actual number of divisions in the clone giving rise to tree t, we generated synthetic data sets where all the clones had the same fixed number of generations. Multiple data sets were created spanning the range from 5 through 9 generations. For each of these data sets, the mutation rate was estimated (assuming the actual value for $\lambda$ is known), and the average number of divisions among all the clones in the data set was predicted. This prediction was positively correlated with the actual number of generations (Figure 8a). It was accurate when the number of divisions was high. When the number of divisions was below ~7 this method overestimates the average number of divisions, which leads to an underestimate in the predicted mutation rate for these clones. This linkage makes intuitive sense since a clone that has longer to mutate can achieve the same frequency of mutations with a lower mutation rate. The direction of causality is suggested by the observation that the method overestimates the number of divisions at low generation numbers even when the mutation rate $\mu$ is known (Figure 8b). In these cases our method provides a lower bound on the mutation rate.

## 3.4 Analysis of autoimmune data

We applied our methods to estimate mutation parameters from a set of experimentally derived lineage trees collected from autoimmune

mice (William *et al*., 2002). Details of the tree shapes are described in (Kleinstein *et al*., 2003) and (Magori-Cohen *et al*., 2006). From these data we estimate using the simulation-based method that ~55% of FWR replacement mutations are lethal (i.e., $\lambda = 0.55$), that the clones have undergone ~5 divisions on average, and that the mutation rate is $\geq 0.26$ generation$^{-1}$, corresponding to approximately $0.26/340 = 7.6\text{x}10^{-4}$ base-pair$^{-1}$ generation$^{-1}$. The analytical method produces a similar estimate of the mutation rate. We can conclude that these clones are undergoing hypermutation at a rate consistent with a 'classic' immune response (McKean *et al*., 1984; Wabl *et al*., 1985; Wabl *et al*., 1987). This is an important (and surprising) result since these cells were microdissected from the T zone-red pulp border rather than germinal centers, where hypermutation is thought to be restricted. This has significant implications for understanding the etiology of autoimmune diseases such as Lupus.

## 4 SUMMARY

Estimating mutation properties is a key element in much of the theory dealing with evolution of cells or species. Currently existing methods assume a very high number of generations, and often a large population size. These assumptions do not apply in many systems, such as the short-term evolution of viruses in a human host or B cell affinity maturation during an immune response (the main focus of this paper). We have developed a MBDP simulation to model the B cell affinity maturation process, along with two ML methods for estimating the mutation parameters (including somatic hypermutation rate, lethal mutation frequency and the number of generations). The input to our methods consists of a set of maximum parsimony lineage trees generated from experimentally observed groups of clonally related B cell receptor DNA sequences. The correctness of the maximum parsimony reconstructions was tested on synthetic data sets and found to be precise for over 98% of trees. Our methods are based on an initial selection of the most informative tree shapes (based on mutual information). In the first method, we derive analytical estimates for the expected value of each tree shape given a set of parameters, and compare these with the observed shapes using weighted least squares. The second method, based on numerical simulations of the underlying MBDP, was developed to cover cases where repeated sequences could be artifacts of the specific experimental protocols employed. Although limited by its high computational requirements, it has the additional advantage of estimating the full joint distribution of tree shapes instead of estimating the expected value of each shape individually. The analytical method can be viewed as a first rapid approximation to this full distribution estimate. The validity of these methods is verified using synthetic data sets. Our methods provide unbiased estimates of the mutation rate, the lethal mutation frequency and the age (in cell generations) of each tree when the number of generation is higher than seven, and a lower bound for younger trees, even for cases where the amount of data is limited.

Preliminary results suggest that our current approach fails to estimate the rate of mutation-independent cell death. We have generated data sets similar to the one used in the analysis above and included mutation-independent cell death with a rate of $\delta$ per division, and attempted to estimate $\delta$. We found that the ML curves were too flat in the direction of $\delta$ to provide any insight (data not shown). This result is not surprising since cell death is equivalent to missing a full branch in the cell-sampling step, which is a frequent occurrence in any case due to the small number of cells sampled to create each tree. Consequently, we expect the MBDP parameter estimates will be unaffected by assuming $\delta = 0$.

It is possible to extend our MBDP model to include other biological processes, such as selection. Negative selection is currently included in the analysis in the form of lethal mutations, but there is currently no positive selection (for a discussion of why this is not critical for the particular experimental data analyzed, see (Kleinstein *et al*., 2003)). Positive selection in its simplest form could be described using a model with two populations with equal mutation rates, but one dividing (or dying) faster than the other. In such a simple model the ratio between the variance in the number of mutations per sequence in the trees and their average would be greater than one. On the other hand, this ratio is not sensitive to the occurrence of lethal mutations and we find in our synthetic data that, as expected, the average is equal to the mean (Magori-Cohen *et al*., 2006). Significant deviations from one would suggest the presence of positive selection, and require that an appropriate model be built for the relevant system. Note that evolutionary relationships (Goldman 1994), or non-homogeneous sampling might also result in ratios higher than one.

We expect that the low frequency of tree reconstruction errors (less than 1.5%) will have a limited effect on the final parameter estimates since these combine multiple structural elements with elements taken purely from the sequences (e.g., R/(R+S)). Another assumption that could impact our methods validity is that of synchronous division. While we expect this assumption will be approximately true of the *in vivo* data as a result of the small microdissections and short time-scales being considered, we have used synthetic data to show that relaxing this assumption does not significantly affect our results (Figure 6 inset).

In summary, we have developed a rapid, systematic measure of mutation parameters from small sets of DNA sequence data, based on a limited set of lineage tree shapes deemed most relevant to the underlying process. We have further provided a methodology based on comparison with synthetic data to test the limits of its applicability.

## ACKNOWLEDGEMENTS

The work of Y.L. and R.M.C. was covered by BSF grant 2003328 and the EU 6th framework co3 pathfinder. S.H.K. was supported in part by NSF IGERT grant DGE-9972930.

## REFERENCES

Clement,M. *et al*. (2000) ''TCS: a computer program to estimate gene genealogies.'' *Mol. Ecol*., **9**(10), 1657–9.

Dunn-Walters,D.K. *et al*. (2002) ''The dynamics of germinal centre selection as measured by graph-theoretical analysis of mutational lineage trees.'' *Dev. Immunol*., **9**(4), 233–43.

Dunn-Walters,D.K. *et al*. (2004) ''Immune system learning and memory quantified by graphical analysis of B-lymphocyte phylogenetic trees.'' *Biosystems*., **76**(13), 141–55.

Goldman,N. (1994) ''Variance to mean ratio, R(t), for poisson processes on phylogenetic trees.'' *Mol. Phylogenet. Evol*., **3**(3), 230–9.

Kleinstein,S.H. *et al*. (2003) ''Estimating hypermutation rates from clonal tree data.'' *J. Immunol*., **171**(9), 4639–49.

Kleinstein,S.H. and Singh,J.P. (2001) ''Toward quantitative simulation of germinal center dynamics: biological and modeling insights from experimental validation.'' *J. Theor. Biol.*, **211**(3), 253–75.

Kleinstein,S.H. and Singh,J.P. (2003) ''Why are there so few key mutant clones? The influence of stochastic selection and blocking on affinity maturation in the germinal center.'' *Int Immunol.*, **15**(7), 871–84.

Magori-Cohen,R. *et al.* (2006) , http://www.cs.princeton.edu/~stevenk/trees.

McKean,D. *et al.* (1984) ''Generation of antibody diversity in the immune response of BALB/c mice to influenza virus hemagglutinin.'' *Proc. Natl Acad Sci. USA*, **81**(10), 3180–4.

Mehr,R. *et al.* (2004) ''Analysis of mutational lineage trees from sites of primary and secondary Ig gene diversification in rabbits and chickens.'' *J. Immunol.*, **172**(8), 4790–6.

Nee,S. *et al.* (1994) ''Extinction rates can be estimated from molecular phylogenies.'' *Philos. Trans. R Soc. Lond. B Biol. Sci.*, **344**(1307), 77–82.

Rosenberg,N.A. and Nordborg,M. (2002) ''Genealogical trees, coalescent theory and the analysis of genetic polymorphisms.'' *Nat. Rev. Genet.*, **3**(5), 380–90.

Shlomchik,M.J., Litwin,S. and Weigert,M.G. (1990) The influence of somatic hypermutation on clonal expansion. *rogress in Immunology.* In *Proceedings of the Seventh International Congress of Immunology,* **7**, 415.

Shlomchik,M.J. *et al.* (1998) ''Clone: a Monte-Carlo computer simulation of B cell clonal expansion, somatic mutation, and antigen-driven selection.'' *Curr. Top Microbiol. Immunol.*, **229**, 173–97.

Takezaki,N. *et al.* (1995) ''Phylogenetic test of the molecular clock and linearized trees.'' *Mol. Biol. Evol.*, **12**(5), 823–33.

Wabl,M. *et al.* (1985) ''Hypermutation at the immunoglobulin heavy chain locus in a pre-B-cell line.'' *Proc. Natl. Acad Sci. USA*, **82**(2), 479–82.

Wabl,M. *et al.* (1987) ''Measurements of mutation rates in B lymphocytes.'' *Immunol Rev.*, **96**, 91–107.

Wagner,S.D. and Neuberger,M.S. (1996) ''Somatic hypermutation of immunoglobulin genes.'' *Annu. Rev. Immunol.*, **14**, 441–57.

William,J. *et al.* (2002) ''Evolution of autoantibody responses via somatic hypermutation outside of germinal centers.'' *Science*, **297**(5589), 2066–70.