# *Inferring complex substitution dynamics given flanking nucleotides, G+C content and recombination rate*

*Amos Tanay, Department of Computer Science and Applied Mathematics, The Weizmann Institute.*
76100, Rehovot, Israel.

Phone: 972-8-9343579
Email: amos.tanay@weizmann.ac.il

## ABSTRACT

Computational models for sequence evolution are a corner stone of molecular evolution. With the rapidly increasing availability of genome sequences it is becoming clear that even the neutral evolutionary process is a complex and heterogeneous one. Subsequently, models for neutral evolution must consider the distinct evolutionary regimes in different genomic regions and model accurately context dependent substitution dynamics. This is of particular importance in comparative genomics applications, where a neutral background is routinely assumed to test for sequence conservation and assign putative function. The model presented here provides a general framework for studying substitution dynamics given complex sequence contexts. It is based on a probabilistic formulation that can express multiple types of context effects and provides efficient algorithmic solutions for inference of ancestral genomes and estimation of model parameters. Importantly, the model and its implementation are adequate for modeling whole genomes, thereby fully exploiting the vast amount of available data to learn parameter-rich models robustly. When applied to almost 10 billion nucleotides of primate sequences, the model construct the most comprehensive characterization of neutral substitution dynamics to date. The results reflect surprising connections between the rate of different mutations types, flanking nucleotides and regional G+C contents. Comparison of the substitution dynamics in apes and monkeys further demonstrate a dynamic and lineage specific neutral process. These results indicate that rich models for the evolution of sequences without a functional constraint are feasible and may allow more accurate comparative genomics for detecting selection on complex sequence potentials.

## INTRODUCTION

More than forty years ago, molecular evolution emerged as a revolutionary paradigm that transformed evolutionary theory and biology. The pioneering works of Kimura and his peers established new principles in times when genomes where still very poorly characterized and when sequence data was scarce. The mathematics supporting molecular evolution was therefore aiming at an economical representation of the evolutionary process using key parameters. With the technological and experimental advances of recent years, genome sequences have become widely available, making parameter rich models for molecular evolution feasible. The comparisons of whole genomes have started to reveal a heterogeneous and dynamic evolutionary process and outlined numerous correlations between genomic features and the evolutionary dynamics of their surroundings. Consequently, comprehensive evolutionary analysis of the relations between the many effectors and consequences of the evolutionary process remains challenging and promises to be rewarding.

Several authors have responded to these challenges by developing new models for neutral sequence evolution[1-5]. These models are based on the understanding that the flux of mutations (the *mutational input*) at a given locus is strongly affected by the immediate (e.g., flanking nucleotides) and regional (e.g. cross over rate) context of that locus. The new models proposed are introducing large sets of parameters and considerable complexity, making computation and expandability important issues. The correlation between so many genomic features and neutral evolution rates poses questions on cause and effect, and make the interpretation of inferred model parameters non trivial.

In this work we introduce a new flexible and parameter rich model for sequence evolution. The model follow up on current context-aware evolutionary models [1, 4, 5] while developing a computational infrastructure for further expansion of the paradigm. Our new model is generalizing the notion of context and developing algorithms that scale well with the massive challenge of inferring parameters and ancestral sequences from whole genome alignments [6]. We applied our method to almost $10^{10}$ bases of human, chimpanzee and rhesus macaque sequence and derived statistically robust parameters for describing neutral evolution in primates. Analysis of the new model puts several theories and analyses that focused on the correlation between specific evolutionary effectors on common grounds [7-9]. Furthermore, our analysis reveals unexpected new connections between flanking context, G+C content and recombination rates and shows that lineage specific changes in the mutational input are more extensive than previously thought. A rich yet robust model for neutral evolution is therefore critical for deriving accurate and informative comparative genomics.

## METHODS

**A flexible context-model for sequence evolution.** We model a group of aligned genomic sequences over a known phylogeny using a graphical model that generalizes existing models for sequence evolution. Here we define the model formally in its basic form, postponing interpretations and comparisons to other models to the discussion. The model is a factor graph [10] defining a joint distribution over random variables of three types: sequence variables, context variables and regional variables (**Fig 1**). We use the common notational convention, denoting random variables by capital letters, assignment to random variables by lower case letters, and assignment to groups of random variables by bold face lower case letters.

-We denote by $S_i^j$ the *sequence variable* representing the nucleotide at locus *j* in the genome of species *i*. Sequence variables are introduced for both extant and ancestral species, according to the assumed phylogenetic tree T. We use *pa(i)* to denote the parent of species *i* and *r* do denote the root of the phylogeny.

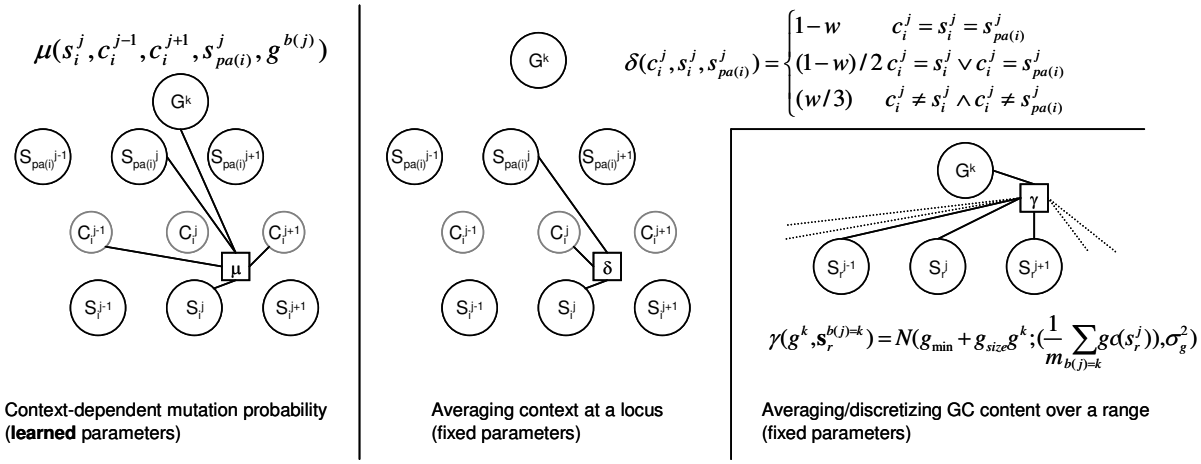


**Figure 1: The Mutation, context and G+C factors.** Shown are the three main factor types in our model, see text for details.

-We denote by $C_i^j$ the *context variable* representing the distribution of nucleotides at locus *j* during evolution over the lineage leading to species *i* from *pa(i)* (see discussion for interpretations).

- We denote by $G^k$ the *regional GC variable* representing the discretized GC content in a range of sequence variables *[km..(k+1)m)*, where *m* is the size of the range. We denote by b(i) the mapping between loci and ranges (i.e., int(i/m)). $G^k$ can be assigned with B possible values, each representing a fixed interval of G+C percentages. Other regional genomic and evolutionary properties can be expressed using similar variables.

Given the above random variables, we define the model using *factors* that assign potentials to combinations of variable values. Note that although we define and use these potentials as expressing local conditional probabilities, the model as a whole is undirected, and the global joint distribution it induces is not compatible with the local conditional probabilities defined by the individual factors.

- The mutation factor $\mu(s_i^j, c_i^{j-1}, c_i^{j+1}, s_{pa(i)}^j, g^{b(j)})$ represents the conditional probability of observing a nucleotide at loci j in species *i* given the nucleotide at the same locus in the ancestral species *pa(i)*, the value of the flanking nucleotides during evolution over the lineage and the regional G+C content. A mutation factor is therefore defined by 4x4xB matrices of 4 by 4 elements (each with 12 free parameters). We use distinct matrices for each lineage. We may further divide the genome into regions with different evolutionary parameterization to express more than one evolutionary regime in a single lineage.

- The background factor $\beta(s_r^j, s_r^{j-1}, s_r^{j-2}, g^{b(j)})$ represents the conditional probability of observing a nucleotide at loci j of the root species, given the preceding two nucleotides.

- The context factor $\delta(c_i^j, s_i^j, s_{pa(i)}^j)$ represents the conditional probability of the context variable at locus j and the lineage *i* given the sequence variables at position *j* and the end points of that lineage. δ in the present work is simply averaging the value of the two sequence variables (**Fig 1**).

- The GC factor $\gamma(g^k, \mathbf{s}_r^{b(j)=k}) = N(g_{\min} + g_{size} g^k; (\frac{1}{m} \sum_{b(j)=k} gc(s_r^j)), \sigma_g^2)$ scores the difference between the G+C variable and the mean G+C content in the associated sequence variables using a normal density function with a predefined standard deviation $\sigma_g$. We use G+C content bins of size $g_{size}$ starting from $g_{min}$. *gc(s)* is an indicator function equals 1 for C or G and 0 otherwise.

The joint probability is defined by combining the factor potentials:

$$P(\mathbf{s},\mathbf{c},\mathbf{g}) = \frac{1}{\mathbf{Z}}\prod_j \beta(s_r^j, s_r^{j-1}, c_r^{j-2}, g^{b(j)}) \prod_{i,j} \mu(s_i^j, c_i^{j-1}, c_i^{j+1}, s_{pa(i)}^j, g^{b(j)}) \prod_{i,j} \delta(c_i^j, s_i^j, s_{pa(i)}^j) \prod_k \gamma(g^k, \mathbf{s}_r^{b(i)=k})$$

where Z is a normalization factor (also known as the partition function).

**Inference.** Given a model (the parameterization of all factors) and a multiple alignment (on a set of extant species genomes), we wish to compute posterior distributions of hidden variables (ancestral sequences, context variables) and groups of hidden variables (joint posterior distribution of the variables adjacent to each factor). The posterior distributions of many of the hidden variables in the model are of interest for themselves and can provide us with estimation of ancestral genomes and their properties. More technically, we need to solve the inference problem in order to learn the model parameters (working in an iterative generalized EM algorithm). Since we are aiming at the application of the model to whole genomes, our algorithms must be capable of handling billions of random variables. There are several common approaches for performing inference with complex graphical models and many of them are used in evolutionary contexts. Sampling based inference is relatively simple and accurate but may require significant computational resources. A different class of methods that can be much more efficient is using a variational principle to derive to an approximate representation of the posterior distribution using minimization of a free energy expression (for a gentle introduction with an evolutionary perspective see Jojic et al. [11]). An effective relaxation of the variational approximation leads to a relaxed free energy expression (called the Bethe free energy) and the efficient loopy belief propagation algorithm (LBP [12]). LBP is a message passing algorithm, computing messages from variables to factors and factors to variables until convergence and then using the converged messages to approximate posteriors (see Yedidia et al. [12] for an excellent introduction). Applying the algorithm in its simple form to our model is completely impractical, almost always ending up in non-converging or very slowly converging messages. We are therefore using several algorithmic modifications to make the algorithm feasible and improve its accuracy:

*Initialization*: We are initializing all variables to factor messages by computing a rough estimation for all variable beliefs. This is done by building an approximated locus-independent model for each locus and solving the exact inference problem on that tree (using an analog of the standard Felsenstein's algorithm [13]). Note that this initialization phase is somewhat similar to a single simplified iteration of the structural variational inference technique introduced by Siepel, Jojic and co-authors for the Phylo-HMM model [5, 11]. Our initialization phase is however considering only the free energy terms involving mutation factors for the locus that is being optimized, and is not repeated to convergence.

*Regional messages*: Exact computation of the messages from the GC factor involves summing over an exponential number of assignments to the factor's adjacent variables. Since GC factors have a large number of inputs we must compute approximate messages. To derive the approximation, we observe that because the factor potential depends on the mean G+C content of the associated sequence variables, the original message update rule:

$$m_{\gamma^k \to s_r^j}(s_r^j) = \sum_{g^k, \mathbf{s}_r^{b(j)=k} | s_r^j} \gamma(g^k, \mathbf{s}_r^{b(j)=k}) m_{g^k \to \gamma^k} \prod_s m_{s \to \gamma^k}$$

can be approximated as:

$$m_{\gamma^k \to s_r^j}(s_r^j) \cong \sum_{g^k} \left( \int N(g^k; \nu, \sigma_g^2) \Pr(\nu) d\nu \right)$$

When $\Pr(\nu)$ represents the distribution of the mean GC content given a fixed $s_r^j$ value and the product of variable-to-factor messages from the other sequence variables to $\gamma^k$. The mean is distributed as a sum of

independent indicator variables ($\Pr(gc(s_r^j)=1) = (m_{s_r^j \to \gamma^k}(C) + m_{s_r^j \to \gamma^k}(G))$) and can therefore be approximated using a normal distribution with mean and variance:

$$E_{\Pr(\nu)}(\nu) = (1/m)\left( gc(s_r^i) + \sum_{b(j)=k, j\neq i} \Pr(gc(s_r^j)=1) \right)$$

$$Var_{\Pr(\nu)}(\nu) = (1/(m)) \sum_{b(j)=k, j\neq i} \Pr(gc(s_r^j)=1)(1-\Pr(gc(s_r^j)=1))$$

Approximating the message is done efficiently by numerical approximation of the integral over $\nu$ (in practice, $\nu$'s variance will be very small – making the computation very accurate).

*Super nodes*. We identify loci that were or are likely to include a CpG dinucleotide in current or ancestral genomes and combine them into "super-nodes" when running LBP. The identification is done based on the beliefs we compute during message initialization, using a permissive cut off on the posterior probability for observing a CpG. Running as "super-nodes", the two sequence variables that are potentially associated with a CpG are unified into one random variable, and the related factors are updated accordingly, without changing the overall joint distribution. The messages involving the super node are therefore much larger, and represent accurately the strong coupling between the two loci. This approach is similar to the generalized LBP algorithm, but is simpler to implement since we are eliminating overlaps between super nodes arbitrarily.

**Learning.** We use a generalized EM algorithm that maximizes the Bethe Free energy [12] of the model given the data by running LBP inference and re-estimating the parameters of the mutation and background factors using the inferred joint posterior distribution. Learning model parameters is greatly simplified by assuming each lineage has its own set of parameters (e.g. its own mutation factors parameterization). This assumption makes the maximization step in the generalized EM algorithm straightforward since we do not have to optimize a single rate matrix given several matrix exponentials, or fine tune branch lengths. We note that maximizing the Bethe free energy is not guaranteed to improve the likelihood of the data, but that in practice, using the enhanced inference approach outlined above the algorithm performs robustly. The overall learning of a model using whole genome alignment (3 billion loci over 3-8 species) is a heavy computational task. Our implementation allows massive parallelization over computer clusters by performing inference on 1MB-10kb genomic segments and combining the sufficient statistics from all such segments to complete EM iterations. The results reported here were generated on a 100 cores cluster within hours.

**Data sources and setup.** Sequence alignments where downloaded from the UCSC genome browser. We used UCSC known gene annotation to identify exons, introns and intergenic regions [14]. Recombination data was downloaded from the Hapmap web site [15]. We used the human genome as the reference, and processed all intergenic genomic loci that could be aligned to both chimpanzee and macaque sequences (for the whole genome results), or to at least 4 of the five genomes in the ENCODE analysis. We treated short gaps as missing data. In the distributed running mode, we cut the genome into parts arbitrarily, losing some accuracy on the boundaries. We always ignored mutational parameters for the lineages leading from the root, since these are underdetermined.

## RESULTS

**Estimating a mutational context model.** We developed a new framework for ancestral inference and evolutionary model learning. Our algorithm estimates context dependent substitution probabilities by considering simultaneously the correlations between substitution of different types, flanking nucleotides and regional sequence properties like the G+C content. Given the vast genomic resources available,

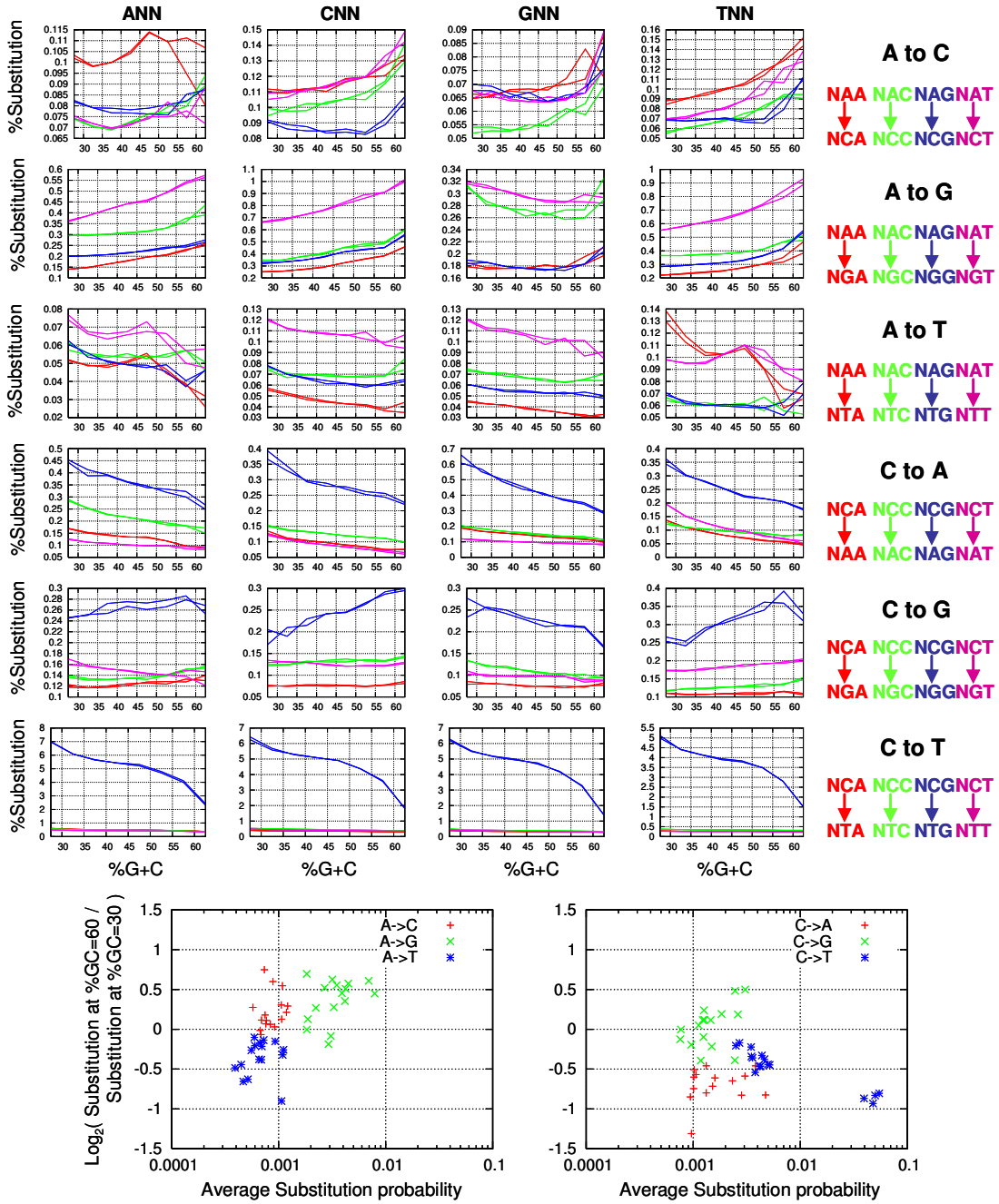


**Figure 2: Parameter rich model for substitution probabilities in the human lineage.** A) Context dependent substitutions. Shown are substitution probabilities of 16 flanking context and 6 mutation types, plotted as a function of G+C content. For each context we plot two independent estimations of the parameters (in two reverse complementing contexts). B) Comparison of substitution probabilities in low (30%) and high (60%) G+C content. Y axis – log of the ratio between substitution probabilities. Positive values reflect substitutions that occur more frequently in region with high G+C content.

learning a parameter rich evolutionary model is statistically robust and since we assume very little on the relations between the different evolutionary parameters, the results provide us with an unbiased view on their interactions. We applied our models to a whole genome alignment of the human, chimpanzee and macaque genomes, processing around 2.8 billion loci (Methods) and to the ENCODE primate sequences

(approximately 30M loci over a 5 species phylogeny). To validate the inferred parameters we used multiple controls. We first verified that our algorithm correctly recover a context-dependent substitution model in simulations (data not shown). We verified that the results are not significantly affected by the quality of the alignment and sequence and that we get similar results when processing repeat-masked and unique sequences. We studied the robustness of the estimated parameters by comparing models for different chromosomes or parts of chromosomes. To illustrate the statistical robustness of the parameters, we will use below comparison of reverse-complementing substitutions in reverse complementing contexts (e.g., ACC to AAC will be matched with GGT to GTT). The difference between such pairs of parameters is used as a quick indication of the expected variance of the parameters, since we learn the reverse complementing parameters independently but expect them to be equivalent (all strand asymmetries[16] should be averaged out since we always consider the plus strand).

**Flanking context and Regional G+C content combine to affect substitution probabilities. Figure 2** depicts the human lineage substitution probabilities for 16 flanking contexts (colors and columns), 8 genomic G+C content ranges (X axis) and the 6 types of mutations (rows). Careful examination of this parameter rich model uncovers surprising correlation trends. The first immediate observation is that the variation in substitution probabilities given flanking contexts is considerable. This is true even if we ignore the well documented variation in CpG deamination (C to T mutations in NCG contexts [9]). For loci with the same regional G+C content we observe 3-fold variation in the divergence of A to C and A to T depending on the flanking context, and 4-fold variation in the divergence probabilities of A to G. For mutations involving a C, we see predominantly rapid divergence in contexts involving a CpG (blue curves), not only for the deamination substitution (C to T) but also for C to A and C to G. Context dependent variation in the substitution of C nucleotides is also observed without involvement of a CpG (e.g. over 2 fold increase in CCA to CAA substitutions relative to CCT to CAT substitutions).

A second clear trend is the correlation between the regional G+C content and the substitution probabilities. This trend is markedly different among the different mutation types. As shown in **Fig. 2B**, mutations involving loss of an A and a gain of C or G are positively correlated with the G+C content, while mutations involving a loss of C and gain of A or T are negatively correlated with the G+C content. Interestingly, negative G+C correlation is also observed for substitutions that change an A with a T, even though such substitutions are not changing the G+C content. Even more surprising are the dependencies between flanking context and the G+C effect. For A to G transitions, we observe clear G+C correlation when the flanking context does not include a G before the mutated locus, but no such correlation when a G exists before the locus. These surprising effects are adding to the theoretical challenge of explaining the origin of isochore structure and G+C heterogeneity in the genome. For example, substantial evidence suggest that increase in genomic G+C content may be a consequence of biased gene conversion (BGC [17-19]) which is driven by a preference for G-C pairs over A-T pairs. However, the correlation between G+C content and the rate A to T mutation, and the elimination of G+C content effect on A to G transitions in the presence of a 3' G, (but not a C) are not directly accommodated by this theory.

**The substitution spectrum in recombination hotspots.** One of the key factors hypothesized to shape the genome's nucleotide composition is the variability in recombination rate. We therefore computed context-dependent substitution parameters in recombination hot-spots (derived from high resolution Hapmap data) and compared them to the global substitution parameters (**Fig 3**). The results indicate that the substitution spectrum is strongly correlated with the recombination process. For example, transitions of an A to a G are occurring more rapidly in recombination hot spots, regardless of the flanking context or the regional G+C content (**Fig 3**). The symmetric process, however, substituting G's with A's is occurring at similar rates in recombination hot spots and other parts of the genome. The same asymmetry in the effect of recombination is observed when analyzing substitutions of A's with C's (faster substitutions of A with C in recombination hot spots).

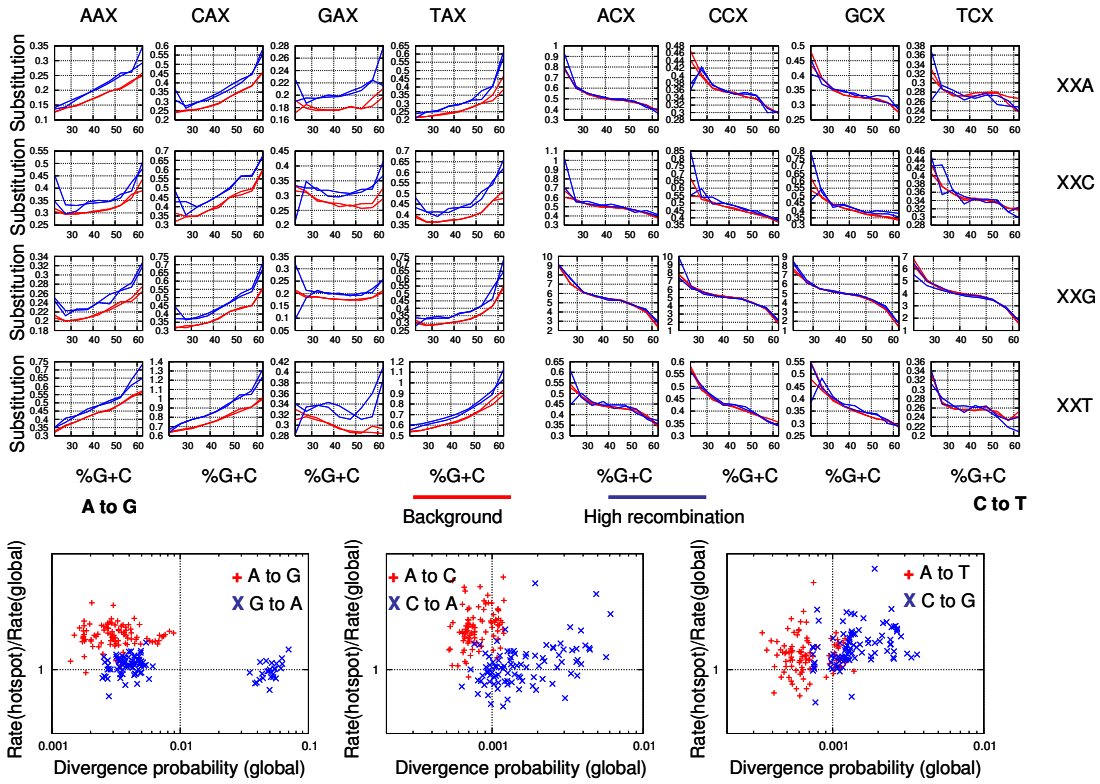


**Figure 3: The effect of recombination on context specific substitutions.** Shown are context dependent A to G (upper left) and C to T (upper right) substitution probabilities as a function of the G+C regional content (X axis). Plotted in red are the global genomic substitution probabilities, and in blue are those computed using only sequences from recombination hotspots. The ratios between the recombination- dependent and non recombination-dependent substitution probability for all six mutation types is shown in the lower panels. Each point in the plots represents comparison of one parameter (fixing flanking context and G+C content).

Recombination rate and G+C content are known to be highly correlated and several theories were put forward to try and explain this correlation. Our data suggest that the correlation between G+C content and substitution rates is significant even when restricting the analysis to recombination hotspots (**Fig. 3A**). Moreover, the amount of recombination-dependent increase in substitutions gaining a G or C is not significantly dependent on the regional G+C content. We also do not observe correlation between the flanking context and the recombination-dependent increase in substitution rate (in contrast to the correlation between flanking context and the G+C content influence on substitutions, **Fig 2**). We conclude that recombination is clearly a factor that can be associated with increase in G+C content, but that it is difficult to argue that recombination rates can explain the correlation between G+C content and substitution dynamics. We note that refining the analysis to include different ranges of cross over rates (beyond the classification to "hot spots" and "background") did not reveal significant additional correlation between recombination and substitution rates (data not shown).

**Context-dependent substitution spectrum in apes and old-world monkeys.** The context-dependent substitution spectra of the lineages following divergence of the apes from the old-world monkeys were estimated using ENCODE-derived sequences from baboon, macaque, chimpanzee and human, with marmoset serving as an out-group. Comparing substitution probabilities between lineages should be carefully approached, as the estimation of substitution rate in closely related species may be affected by demographic factors (which should affect the entire spectrum) or sequencing errors (which can be

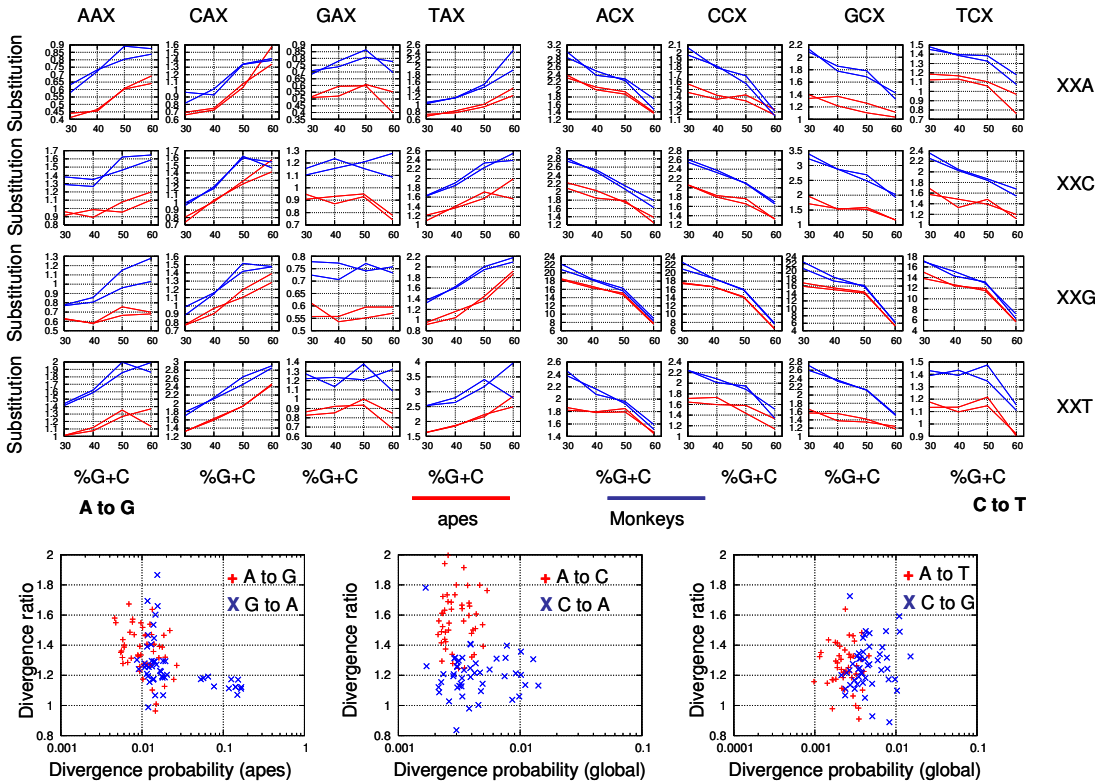


**Figure 4: substitution dynamics in apes and monkeys.** Shown are context dependent A to G (upper left) and C to T (upper right) substitution probabilities as a function of the G+C regional content (X axis). Plotted in red are the parameters from the ape lineage (leading to the human-chimp common ancestor) and in blue the parameters for the monkey lineage (leading to the baboon-macaque common ancestor). The ratios between the parameters in the two lineages for all six mutation types are shown in the lower panels. Each point in the plots represents comparison of one parameter (fixing flanking context and G+C content).

context- and lineage-specific). For example, error rates in the first chimpanzee assembly (panTro1) are not negligible when compared to the human chimp divergence probabilities (Data not shown). The comparison reported here is using internal phylogenetic branches that are supported by at least two species on each side, providing robust results which are unlikely to be affected by differences in sequence quality. **Fig 4A,B** depicts the differences in substitution probabilities between apes and monkeys. The same general trends we observed in the genomewide analysis of the human lineage (e.g., G+C dependencies, cross talk between flanking nucleotide and G+C correlations) are observed in both lineages, but significant differences between lineages are evident. Comparison of CpG and non CpG substitution rate was suggested before to indicate that two molecular clocks are ticking in the apes and monkeys lineages [20]. A generation-time based clock was suggested to be coupled to non CpG substitutions and an absolute time clock was suggested to be associated with CpG deamination. The generational clock was estimated to tick 1.36 (log ratio = 0.44) faster in the monkeys' lineage. As shown here, several types and substitutions (e.g., A to C substitutions) are accelerated by up to twice that rate (log ratio > 0.8), while other substitution types (C to A, C to G) are occurring at almost the same rates in the apes' and monkeys' lineages. Consequently, no single molecular clock, and not even two distinct molecular clocks can describe the evolutionary process in apes and monkeys. One possibility is that for each substitution type and each context, a different combination of the generational and absolute time clocks is determining the substitution rate. Another possibility is that on top of these effects, small changes in the mutational input and output contribute another layer of complexity**.**

## DISUSSION

**Models and approximations for context dependent evolutionary processes**. If we assume that neutral evolution progresses through a series of substitutions, and ignore the population dynamics so as to assume that mutations and fixations are instantaneous, then we can think of the entire (neutral) evolutionary process as determined by the rate of substitution at each locus given the current sequence (or the *context*). If the context is void (i.e., the mutational input at each locus is independent of the other loci), then the process can be described as a product of independent continuous time Markov processes, each determined by some stationary (fixed throughout the process) rate matrix. In this simple case, one can easily compute the likelihood of a set of sequences over a phylogeny, by multiplying the likelihoods on individual loci, where the substitution probabilities on each lineage are derived using exponentiation of the rate matrix with appropriate branch lengths. It is important to note that the matrix exponential can be thought of as marginalizing (or integrating) the probabilities over all possible evolutionary trajectories with fixed nucleotide values at a parent and child species. The context-independence assumption therefore makes it possible to marginalize over all possible evolutionary trajectories of a long sequence by considering one locus at a time.

In the case where the context is not void we can no longer decompose the probability of an evolutionary trajectory to independent contributions from individual loci, and we can therefore no longer marginalize over all trajectories by solving small matrix exponentials and multiplying them together. Even when the context include only the two nucleotides flanking a loci, the simultaneous change of the entire sequence make it impossible to represent the likelihood as a product of independent terms. Computing exact likelihoods in this case is therefore highly intractable (the dimension of the trajectory space is truly immense) and one needs to find effective approximations or heuristics. One way by which the problem can be simplified is through time discretization. Here, the time domain is considered as a set of discrete intervals (lineages in the phylogenetic tree [5] or subdivision of lineages to smaller time intervals [4, 21]. The Markov process at each locus is now assumed stationary between each two time points. For example, one can assume the context of the process from time t to time t+1 to be determined completely by the sequence at time t [4, 7], or to be determined by the combination of values at time t and t+1 [11]. Under this assumption, the likelihood of the data become much easier to understand and compute using standard statistical methods. Even then, the problem is computationally hard and requires additional approximation. For example, Siepel and Haussler have further simplified the process to include flanking effects at only one side of the locus, and developed a Bayesian network model for expressing the joint distribution of extant and ancestral sequence variables [5] and variational inference methods to allow efficient model learning . Their influential work became an important foundation for popular comparative genomics tools [22]. Hwang and Green [4] have introduced a time synchronous model that include flanking effects from the 3' and 5' nucleotides of each loci and used the Markov Chain Monte Carlo method to estimate the model parameters. Arndt and his colleagues used a similar methodology but also developed an intelligent framework to identify statistically significant context rules when the data is limiting [1, 2]. Duret and Arndt also introduced methodologies for special treatment of the important case of CpG dinucleotides and explored the connections between G+C content and recombination rates [7].

The discrete and synchronous time models we described above generally fall under the umbrella of the dynamic Bayesian Network formalism. A more sophisticated class of computational models for studying context dependent Markov processes is recently being developed by Koller and colleagues. Noodleman and Koller introduced the Continuous Time Bayesian Network (CTBN) paradigm, which treats a set of time-evolving random variables that are parameterized by context dependent Markov rate matrices [23], where the context dependencies are represented by an arbitrary graph. To perform inference in the model, Noodleman and co-workers developed a series of message passing algorithms [24] which exchange beliefs on rate matrices (instead of beliefs on stationary distributions as commonly practiced with static Bayesian networks). To make this approach practical, even when the models are not very large, one still have to assume independence among the messages in the model, but here, in contrast to the synchronous

time approximations, the algorithm develops some characterization of the posterior distribution over trajectories (approximated by independent Markov processes at each locus), and not only information on the endpoints distribution. El-hay and colleagues [25] have described an evolutionary-inspired variant to the CTBN framework. In their Continuous Time Markov Network model, the process is parameterized by a context independent reversible Markov process that continuously propose changes to variables. This proposal mechanism is combined with a Markov network that determine stochastically if proposed changes are accepted based on the context. This formulation can be thought of as describing evolution given a mutational input (the proposal distribution) and fitness function (the Markov network).

In the present work, we introduced a model that extends the time-synchronous dynamic Bayesian networks using simple context variables that represent the distribution of sequence variables during evolution over a lineage. By adding these layers of variables, our model develops information on the evolutionary trajectories at each locus as part of the inference procedure. This is done without having to work with the large and more complex messages that are required by the CTBN framework. This (relative) simplicity is important since we are working with huge models, and inference performance is a major consideration. The message passing algorithm we have developed is flexible enough to allow both very fast implementation and specific handling of important cases like strongly coupled CpG dinucleotides. In learning the model, we have assumed that sequence data is not a major limiting factor, and since our observations suggest that the context–dependent Markov process is far from being stationary or reversible, even when considering lineages as close as human and chimp, we opted for the usage of lineage-specific parameters and avoided altogether the explicit estimation of rate matrices. We believe the model is therefore representing a cost-effective tradeoff between complexity and efficiency, which will be exceedingly important as we continue to build into it additional layers to represent selection.

**G+C content, recombination and the mutation spectrum.** The model we introduced here is not assuming any parametric a-priori dependency between substitution probabilities and other genomic features. It can therefore be used to examine parametric hypotheses on the nature of correlation between genomic features and substitution dynamics in an unbiased way. An important example is G+C genomic heterogeneity. Our data reveal extensive correlation between higher rate of G and C gaining substitutions and higher G+C content. This correlation was observed before and is subject to extensive theoretical and computational work, trying to rationalize the genome isochore structure as a result of current evolutionary dynamics or as a historical leftover of ancestral events (few examples are in [8, 18, 19]). Many of these theories are predicted to affect several mutation types, or several flanking context in the same way, generating hypotheses that can be tested with our model. For example, our data show that the G+C content effect on substitution dynamics is holding even when restricting the analysis to HapMap recombination hotspots (**Fig 3**). Even more intriguingly, the strong G+C effect on A to G transitions is eliminated for loci with a 5' G nucleotide (**Fig 2**). Theories that rely on the recombination structure of the genome to explain G+C heterogeneity should somehow accommodate for these new observations.

**Application to comparative genomics**. The model presented here as well as other recent evidence [20, 26] shows that the genomic substitution process is very dynamic among lineages and cannot be accurately modeled using a universal rate matrix, even if this matrix is expanded to include context effects. Universal substitution models are still critically important for studying more remote species or when using likelihood based method for phylogenetic reconstruction. In the current typical comparative genomics settings, where ample of sequence data is available and the universality of the process is questionable, it makes little sense to assume parametric dependencies among the substitution probabilities in different lineages. The rich and unbiased models we derive here can greatly enhance the accuracy of the neutral model we use when searching for selected sequence traits. Such enhanced neutral standard would allow evolutionary biologists to develop better tools for detecting selection on complex sequence potentials like weak transcription factor binding [27] or the nucleosome positioning signals [28].

**REFEFERENCES**


1. Arndt, P.F., C.B. Burge, and T. Hwa, *DNA sequence evolution with neighbor-dependent mutation.* J Comput Biol, 2003. **10**(3-4): p. 313-22.
2. Arndt, P.F. and T. Hwa, *Identification and measurement of neighbor-dependent nucleotide substitution processes.* Bioinformatics, 2005. **21**(10): p. 2322-8.
3. Arndt, P.F., T. Hwa, and D.A. Petrov, *Substantial regional variation in substitution rates in the human genome: importance of GC content, gene density, and telomere-specific effects.* J Mol Evol, 2005. **60**(6): p. 748-63.
4. Hwang, D.G. and P. Green, *Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution.* Proc Natl Acad Sci U S A, 2004. **101**(39): p. 13994-4001.
5. Siepel, A. and D. Haussler, *Phylogenetic estimation of context-dependent substitution rates by maximum likelihood.* Mol Biol Evol, 2004. **21**(3): p. 468-88.
6. Ma, J., et al., *Reconstructing contiguous regions of an ancestral genome.* Genome Res, 2006. **16**(12): p. 1557-65.
7. Duret, L. and P.F. Arndt, *The impact of recombination on nucleotide substitutions in the human genome.* PLoS Genet, 2008. **4**(5): p. e1000071.
8. Karro, J.E., et al., *Exponential decay of GC content detected by strand-symmetric substitution rates influences the evolution of isochore structure.* Mol Biol Evol, 2008. **25**(2): p. 362-74.
9. Tanay, A., et al., *Hyperconserved CpG domains underlie Polycomb-binding sites.* Proc Natl Acad Sci U S A, 2007. **104**(13): p. 5521-6.
10. Kschischang, F.R., B.J. Frey, and H.-a. Loeliger, *Factor graphs and the sum-product algorithm.* IEEE Transactions on Information Theory, 2001. **47**: p. 498--519.
11. Jojic, V., et al., *Efficient approximations for learning phylogenetic HMM models from data.* Bioinformatics, 2004. **20 Suppl 1**: p. i161-8.
12. Yedidia, J.S., W.T. Freeman, and Y. Weiss, *Constructing free energy approximations and generalized belief propagation algorithms.* IEEE tran info theory, 2005. **51**(7): p. 2282-2312.
13. Felsenstein, J., *Evolutionary trees from DNA sequences: a maximum likelihood approach.* J Mol Evol, 1981. **17**(6): p. 368-76.
14. Rosenbloom, K., et al., *Phylogenomic resources at the UCSC Genome Browser.* Methods Mol Biol, 2008. **422**: p. 133-44.
15. Frazer, K.A., et al., *A second generation human haplotype map of over 3.1 million SNPs.* Nature, 2007. **449**(7164): p. 851-61.
16. Polak, P. and P.F. Arndt, *Transcription induces strand-specific mutations at the 5' end of human genes.* Genome Res, 2008. **18**(8): p. 1216-23.
17. Dreszer, T.R., et al., *Biased clustered substitutions in the human genome: the footprints of male-driven biased gene conversion.* Genome Res, 2007. **17**(10): p. 1420-30.
18. Galtier, N. and L. Duret, *Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution.* Trends Genet, 2007. **23**(6): p. 273-7.
19. Galtier, N., et al., *GC-content evolution in mammalian genomes: the biased gene conversion hypothesis.* Genetics, 2001. **159**(2): p. 907-11.
20. Kim, S.H., et al., *Heterogeneous genomic molecular clocks in primates.* PLoS Genet, 2006. **2**(10): p. e163.

21. Raijman, D., R. Shamir, and A. Tanay, *Evolution and selection in yeast promoters: analyzing the combined effect of diverse transcription factor binding sites.* PLoS Comput Biol, 2008. **4**(1): p. e7.
22. Siepel, A., et al., *Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.* Genome Res, 2005. **15**(8): p. 1034-50.
23. Nodelman, U., C.R. Shelton, and D. Koller. *Continuous Time Bayesian Networks*. in *UAI2002*. 2002. Edmonton, Alberta, Canada: Morgan Kaufmann.
24. Nodelman, U., D. Koller, and C.R. Shelton. *Expectation Propagation for Continuous Time Bayesian Networks*. in *UAI2005*. 2005. Edinburgh, Scotland: AUAI Press.
25. El-Hay, T., et al. *Continuous Time Markov Networks*. in *UAI2006*. 2006. Cambridge, MA, USA: AUAI Press.
26. Tyekucheva, S., et al., *Human-macaque comparisons illuminate variation in neutral substitution rates.* Genome Biol, 2008. **9**(4): p. R76.
27. Tanay, A., *Extensive low-affinity transcriptional interactions in the yeast genome.* Genome Res, 2006. **16**(8): p. 962-72.
28. Segal, E., et al., *A genomic code for nucleosome positioning.* Nature, 2006. **442**(7104): p. 772-8.