Identification and Measurement of Neighbor Dependent Nucleotide Substitution Processes

Peter F. Arndt

Department of Computational Molecular Biology
Max Planck Institute for Molecular Genetics
Ihnestr. 73
14195 Berlin, Germany
arndt@molgen.mpg.de

Abstract: The presence of different neighbor dependent substitution processes generates specific patterns of dinucleotide frequencies in all organisms. Based on a general framework of how to include such processes into more realistic models of nucleotide substitutions we develop a method that is able to identify such processes, measure their strength, and judge their importance to be included into the modeling. Starting from a model for neighbor independent nucleotide substitution we successively add neighbor dependent substitution processes in the order of their ability to increase the likelihood of the model describing the data. The analysis of neighbor dependent nucleotide substitutions in human, zebrafish and fruit fly is presented.

1 Introduction

The identity of the neighboring nucleotide can have a drastic influence on the mutation rates of a nucleotide. A well-known and studied example of this fact is the increased mutation of cytosine to thymine in CpG dinucleotides in vertebrates [Co78, RR80]. This process is triggered by the methylation of cytosine in CpG followed by deamination, and mutation from CpG to TpG or CpA (on the reverse strand). Due to this process the number of CpG is decreased while the number of TpG and CpA is larger than expected from independently evolving nucleotides. Most of the deviant dinucleotide odds ratios (dinucleotide frequencies normalized for the base composition) in the human genome can be explained by the presence of the CpG methylation deamination process [ABH02]. Biochemical studies in the 1970s already compared these odds ratios for different genomes and different fractions of genomic DNA [Ru76, RS77] and concluded that these ratios are a remarkably stable property of genomes. In the following Karlin and coworkers [CB95, KM97, KMC97] elaborated and expanded these observations, showing that the pattern of dinucleotide abundance constitutes a genomic signature in the sense that it stable across different parts of a genome and generally similar between related organisms. This suggests that the causes of these genomic signatures are inheritable. Since this signature is also present in non-coding and intergenic DNA it is very promising to study neighbor dependent mutation and fixation processes (we refer to the effective process as the substitution process) to understand the driving force behind these remarkably property of genomes.

Recently framework to include such neighbor dependent processes has been introduced and successfully applied to model the CpG methylation deamination process [ABH02, APH03]. Here we will extend this model and discuss the inclusion of other neighbor dependent substitutions and how one can infer their relevance without prior knowledge on the underlying biochemical processes. In vertebrates the CpG methylation deamination process is the predominant nucleotide substitution process. Its rate is about 40 times higher than this of a transversion and its history can actually reconstructed for the last 250 Myr [APH03]. One reason for this substitution frequency being so high is that methylation in vertebrates is also used in gene regulation - as a byproduct methylated CpG's often mutate. We know already that also other vertebrates use methylation in the same way but do not know about the quantitative extent their genomes are methylated. The situation is still unclear in other kingdoms of life. Although we clearly see a signature of neighbor dependent substitution processes, we do not know the responsible processes and their rates.

To present our method we study neighbor dependent substitutions in zebrafish (*Danio rerio*) and fruit fly (*Drosophila melanogaster*). In all these studies we first try to model the observed nucleotide substitutions with a model which does not include any neighbor dependent nucleotide substitutions (12 free rate parameters) and then ask the question which neighbor dependent substitution process one would have to include to describe the observed data best. The idea is to capture the most of the observed substitutions by single nucleotide substitutions independent of the neighboring bases and then to include neighbor dependent substitutions one by one to generate a better model with the least number of parameters. Processes are added in the order of their ability to describe the observed data better. Naturally, the addition of any further process (together with another rate parameter) into a model will increase the likelihood of this model (defined below) to describe the observed data. In order not to over-fit the data we use a likelihood ratio test to judge whether the addition of further process is justified. In this respect our approach is rich in contrast to recently presented work in the same direction by Lunter and Hein [LH04] where 49 parameters are used and neighbor independent substitution processes are not considered separately. The strength of our approach is to come up with models with fewer parameters that still capture the essential neighbor dependent nucleotide substitution processes.

The rest of the paper organizes as follows. In the next section we will describe the details of our method. There is no need to implement the described procedure for readers who want to analyze their own sequences, since we are running a public web server at http://evogen.molgen.mpg.de/server/substitution-analysis . One is able to upload pairs of ancestral and daughter sequences and perform the presented analysis. First applications of such an analysis will be presented in the results section.

2 Method

2.1 The substitution model

In total there are 12 distinct neighbor independent substitution processes of single nucleotides by another. Four of them are so-called transitions that interchange a purine with a purine or a pyrimidine with a pyrimidine. The remaining eight processes are the so-called transversions that interchange a purine with a pyrimidine or vice versa. The rates of these processes will be denoted $r_{\alpha\to\beta}$, where $\alpha, \beta \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ denote a nucleotide. On top of these 12 processes we want to consider also neighbor dependent processes of the kind $\kappa\lambda \to \kappa\sigma$ and $\kappa\lambda \to \sigma\lambda$ where the right or left base of a di-nucleotide changes, respectively. There might be several of those processes present in our model, their rates will be denoted by $r_{\kappa\lambda\to\kappa\sigma}$ or $r_{\kappa\lambda\to\sigma\lambda}$. We do not consider processes where both nucleotides of a dinucleotide change at the same time. In vertebrates the most important neighbor dependent process to consider is the substitution of cytosine in CpG resulting in TpG or CpA. Its rate is about 40 times higher than this of a transversion [APH03]. This process is triggered by the methylation and subsequent deamination of cytosine in CpG pairs. It is commonly (and erroneously) assumed that this process only affects CpG dinucleotides. However this is not the case as it has been shown [ABH02].

The model itself is parameterized by the substitution rates and the length of the time span, dt, the respective substitution processes acted upon the sequence, which would in our case be the time between the observation of an ancestral sequence and its daughter sequence, T. We have the freedom to rescale time and measure it in units of T. In this case, the time span is $dt = 1$ and with this choice the substitution rates are equal to the substitution frequencies giving the number of nucleotide substitutions per bp. In the simplest case our model includes neighbor independent processes only and is parameterized by 12 substitution frequencies. For each additional neighbor dependent process we gain an additional parameter. The set of all these substitution frequencies will be denoted by $\{r\}$. The number of parameters can actually be reduced by a factor of two when one considers substitutions along neutrally evolving DNA. In this case we cannot distinguish the two strands of the DNA and therefore the substitution rates are reverse complement symmetric, e.g. the rate for a substitution C→A is equal to the rate for a substitution T→G (in the following we will denote this rate by $r_{\mathtt{C:G}\to\mathtt{A:T}}$).

In order to facilitate the subsequent maximum likelihood analysis we need to compute the probability, $P_{\{r\}}(\cdot\beta\cdot|\alpha_1\alpha_2\alpha_3)$, that the base α_2 flanked by α_1 to the left and by α_3 to the right, changes into the base β for given substitution frequencies $\{r\}$. This probability can easily calculated by numerically solving the time evolution of the probability to find three bases $p(\alpha\beta\gamma; t)$ at time t, which is given by the Master equation and can be written into the following set of differential equations:

$$\frac{\partial}{\partial t} p(\alpha\beta\gamma; t) \quad = \sum_{\epsilon\in\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} [r_{\epsilon\to\alpha}\, p(\epsilon\beta\gamma; t) + r_{\epsilon\to\beta}\, p(\alpha\epsilon\gamma; t) + r_{\epsilon\to\gamma}\, p(\alpha\beta\epsilon; t)]$$

$$
\begin{aligned}
&+ \sum_{\{\kappa\lambda\to\kappa\sigma\}} r_{\kappa\lambda\to\kappa\sigma}\left[\delta_{\kappa\sigma,\alpha\beta}\, p(\kappa\lambda\gamma;t) - \delta_{\kappa\lambda,\alpha\beta}\, p(\alpha\beta\gamma;t)\right] \\
&+ \sum_{\{\kappa\lambda\to\sigma\lambda\}} r_{\kappa\lambda\to\sigma\lambda}\left[\delta_{\sigma\lambda,\beta\gamma}\, p(\alpha\kappa\lambda;t) - \delta_{\kappa\lambda,\beta\gamma}\, p(\alpha\beta\gamma;t)\right] \qquad (1)
\end{aligned}
$$

where the numbers $r_{\alpha\to\alpha}$ are defined by $r_{\alpha\to\alpha} = -\sum_{\beta\neq\alpha} r_{\alpha\to\beta}$ and $\delta_{\alpha\beta,\gamma\delta}$ is the Kronecker delta

$$
\delta_{\alpha\beta,\gamma\delta} = \begin{cases} 1 & \text{if } \alpha = \gamma \text{ and } \beta = \delta \\ 0 & \text{otherwise.} \end{cases} \qquad (2)
$$

The first three terms describe single nucleotide substitutions on the three sites whereas the last 4 terms represent the neighbor dependent processes on the bases (1,2) or (2,3). At $t = 0$ we start with the three bases $\alpha_1\alpha_2\alpha_3$, which is expressed by the initial condition:

$$
p(\alpha\beta\gamma; t = 0) = \begin{cases} 1 & \text{if } (\alpha\beta\gamma) = (\alpha_1\alpha_2\alpha_3) \\ 0 & \text{otherwise.} \end{cases} \qquad (3)
$$

After numerically iterating the above differential equations using the Runge-Kutta algorithm [Pr92] we get the above transition probability by

$$
P_{\{r\}}(\cdot\beta\cdot|\alpha_1\alpha_2\alpha_3) = \sum_{\beta_1\beta_3} p(\beta_1\beta\beta_3; t = 1)\ . \qquad (4)
$$

The above iteration has to be carried out for all possible 64 combinations of bases to get all 256 possible probabilities $P_{\{r\}}(\cdot\beta\cdot|\alpha_1\alpha_2\alpha_3)$.

2.2 Estimation of substitution frequencies

One can estimate all the above mentioned substitution frequencies by comparing a pair of ancestral $\vec{\alpha} = \alpha_1\alpha_2\ldots\alpha_L$ and daughter sequence $\vec{\beta} = \beta_1\beta_2\ldots\beta_L$, in which the daughter sequence represents the state of the ancestral sequence after the substitution processes acted upon it for some time. Such pairs of ancestral and daughter sequences can be obtained in various ways. One very fruitful approach is to take alignments of repetitive sequences, which can be found in various genomes, and to align them with their respective master sequences as the ancestral sequence [APH03]. Such alignments can be easily retrieved from the RepeatMasker (http://www.repeatmasker.org/). The log likelihood that a sequence $\vec{\beta}$ evolved from a master sequence $\vec{\alpha}$ under a given substitution model parameterized by the substitution frequencies $\{r\}$ then given by

$$
\log L_{\{r\}} = \sum_{i=2}^{L-1} \log P_{\{r\}}(\cdot\beta_i\cdot|\alpha_{i-1}\alpha_i\alpha_{i+1})\ . \qquad (5)
$$

To get estimates for the substitution frequencies for given $\vec{\alpha}$ and $\vec{\beta}$ we then have to maximize the above likelihood by adjusting the substitution frequencies. This can easily be done using Powell's method [Pr92] while taking care of boundary conditions [Bo66], i.e. the positivity of the substitution frequencies. Due to the stochasticisity of the mutation process the estimates of substitution frequencies will inaccurate within some limits. Supplying more sequence data to the algorithm minimizes the error in the frequency estimates.

With the inclusion of additional neighbor dependent processes the likelihood of a model $\{r'\}$ will be greater than the one of the original model $\{r\}$. This is true because the models are nested and one has one free parameter more to fit. To test whether the inclusion of a new parameter is justified we employ the likelihood ratio test for nested models. Let $\lambda = L_{\{r\}}/L_{\{r'\}}$ be the likelihood ratio, then $-2\log\lambda$ has an asymptotic chi-square distribution with degrees of freedom equal to the difference in the numbers of free parameters in the two models, which in our case would be one [EG01]. Since we choose the best of the neighbor dependent processes out of the $4 \times 4 \times 3 \times 2 = 96$ possible such processes we require that $-2\log\lambda > 30$ to have significance on the 5% level and respecting the required Bonferroni correction for multiple testing. We confirmed this conservative threshold by simulations using sequences that have been synthetically mutated according to a known model.

3 Results

As a first test we applied the described method to human genomic data. Here we took the copies of the AluSx SINEs that have been found in a genome-wide search of the human genome (release v20.34c.1 at ensembl.org from April 1st, 2004). These elements are assumed to have evolved neutrally and that therefore the substitution process is reverse complement symmetric. Results are presented in Table 1. In the first column of data we

	6 parameter model	7 parameter model	8 parameter model	9 parameter model
A:T→C:G	0.012	0.012	0.011	0.007
A:T→T:A	0.010	0.011	0.011	0.011
C:G→G:C	0.016	0.016	0.012	0.012
C:G→A:T	0.015	0.014	0.014	0.014
A:T→G:C	0.036	0.036	0.036	0.036
C:G→T:A	0.158	0.059	0.060	0.060
CpG→CpA/TpG		0.618	0.627	0.624
CpG→CpC/GpG			0.029	0.029
TpT/ApA→TpG/CpA				0.013
$-2\log\lambda$		$7.7{\cdot}10^6$	$1.3{\cdot}10^5$	$9.6{\cdot}10^4$

Table 1: Estimates for substitution frequencies for nested models of nucleotide substitution in human AluSx repeats. Given are the substitution frequencies per bp in the time span after the insertion of the AluSx repeats into the human genome. In the last row we note the $-2\log\lambda$ where λ is the likelihood ratio of the model and the one with one less parameter in the column to the left.

give estimations for the 6 neighbor independent single nucleotide substitutions. We test all 96 possible extension of this simple substitution model by one additional neighbor dependent substitution process and its reverse complement symmetric process. As expected (and shown in the second column in Table 1) the CpG methylation deamination process (CpG→CpA/TpG) turns out give the best improvement with $-2\log\lambda = 7.7 \cdot 10^6$, which is clearly above the threshold of 30. The substitution frequency of this process is about 45 times higher than that of a transversion. The second process that needs to be included to improve the model most is the substitution of CpG→CpC/GpG ($-2\log\lambda = 1.3 \cdot 10^5$). This is another CpG based process and probably also triggered by the methylation of cytosine. However, the substitution frequency is about 30 times smaller than this of the CpG→CpA/TpG process. The third process is then the substitution TpT/ApA→TpG/CpA ($-2\log\lambda = 9.6 \cdot 10^4$). The instability of the TpT dinucleotide does not come as a surprise here, since two consecutive thymine nucleotides tend to form a thymine photodimer T $<>$ T. This process is one of the major lesions formed in DNA during exposure to UV light [DZC97].

Next we turn to the analysis of the DANA repeats in zebrafish (*Danio rerio*). Results are presented in Table 2. Again we start with a model just comprising single nucleotide transversions and transitions. As observed in human the transitions occur more often than transversions and there is a strong A:T bias in the single nucleotide substitutions. Zebrafish being a vertebrate also utilizes methylation as an additional process to regulate gene expression. As a consequence we observe a higher mutability of the CpG dinucleotide due to the deamination process also in zebrafish. However the substitution frequency for the CpG→CpA/TpG process is in zebrafish only about 8 times higher than this of a transversion suggesting that the degree of methylation is generally lower than in human.

	6 parameter model	7 parameter model	8 parameter model	9 parameter model
A:T→C:G	0.024	0.025	0.026	0.026
A:T→T:A	0.041	0.041	0.041	0.041
C:G→G:C	0.037	0.036	0.036	0.023
C:G→A:T	0.029	0.029	0.028	0.028
A:T→G:C	0.073	0.074	0.046	0.046
C:G→T:A	0.151	0.111	0.105	0.107
CpG→CpA/TpG		0.274	0.331	0.328
CpA/TpG→CpG			0.100	0.097
CpG→CpC/GpG				0.096
$-2\log\lambda$		$2.9{\cdot}10^5$	$1.6{\cdot}10^5$	$1.1{\cdot}10^5$

Table 2: Estimates for substitution frequencies for nested models of nucleotide substitution in DANA repeats from *Danio rerio*.

We also investigated non-vertebrate sequence data. As an example we present here the analysis of the DNAREP1_DM repeat in *Drosophila melanogaster* (Table 3). The case to include neighbor dependent process is in this clearly not as strong as for vertebrate genomes. The values of $-2\log\lambda$ are 3 orders of magnitude smaller but still above threshold for the first 3 processes which are chosen by our procedure to be included into a model for nucleotide substitutions in fly. The first such process is the substitution TpA→TpT/ApA. Although the corresponding substitution frequency is lower than all the single nucleotide transitions and transversions, the dinucleotide frequencies in the stationary state deviate up to 10% from their neutral expectation under a neighbor independent substitution model [ABH02]. Therefore even processes with a small contribution to the overall substitutions have a large influence on the observed patterns of dinucleotide frequencies or genomic signatures and therefore may very well be solely responsible for the generation of such pattern in different species.

	6 parameter model	7 parameter model	8 parameter model	9 parameter model
A:T→C:G	0.038	0.038	0.038	0.038
A:T→T:A	0.052	0.045	0.045	0.045
C:G→G:C	0.034	0.034	0.034	0.034
C:G→A:T	0.074	0.074	0.074	0.074
A:T→G:C	0.052	0.052	0.052	0.047
C:G→T:A	0.108	0.108	0.098	0.098
TpA→TpT/ApA		0.029	0.028	0.028
TpC/GpA→TpT/ApA			0.036	0.035
GpT/ApC→GpC				0.021
$-2\log\lambda$		853	592	40

Table 3: Estimates for substitution frequencies for nested models of nucleotide substitution in DNAREP1_DM transposable element from *Drosophila melanogaster*.

4 Summary and Outlook

We presented a novel procedure to identify the existence and measure the intensity of neighbor dependent substitution processes. We discussed the extension of a model of nucleotide substitutions in human and included more neighbor dependent processes besides the well-known CpG methylation deamination process [ABH02]. We could also show that the CpG methylation deamination is the predominant substitution process in zebrafish, while it does not play a prominent role in fruit fly. We exemplified our method here using sequence data from one particular subfamily of repeats from the three organisms. In the case of the human genome a much more thorough analysis on various families of repeats have been presented in [APH03]. For zebrafish and fruit fly we presented for the first time a possible extensions to a nucleotide substitution model including neighbor dependencies. In the future we will also analyze bacterial substitution processes and their neighbor dependencies. A comparative study of the predominant substitution processes will further broaden our knowledge about the molecular processes that are responsible for mutation and fixation. This will enable us further to explore the evolution of these processes and possible advantages to various organisms.

Acknowledgment We thank Nadia Singh and Dmitri Petrov (Stanford) for kindly providing sequence data on DNAREP1_DM repeat in *Drosophila melanogaster*.

References

[ABH02] Arndt, P. F., Burge, C. B. and Hwa, T. (2002). *DNA Sequence Evolution with Neighbor-Dependent Mutation.* 6th Annual International Conference on Computational Biology RECOMB2002, Washington DC, ACM Press, KK.

[APH03] Arndt, P. F., Petrov, D. A. and Hwa, T. (2003). *Distinct changes of genomic biases in nucleotide substitution at the time of Mammalian radiation.* Mol Biol Evol 20(11): 1887-96.

[Bo66] Box, M. J. (1966). *A Comparison of Several Current Optimization Methods and Use of Transformations in Constrained Problems.* Computer Journal 9(1): 67-77.

[Co78] Coulondre, C., Miller, J. H., Farabaugh, P. J., et al. (1978). *Molecular basis of base substitution hotspots in Escherichia coli.* Nature 274(5673): 775-80.

[DZC97] Douki, T., Zalizniak, T. and Cadet, J. (1997). *Far-UV-induced dimeric photoproducts in short oligonucleotides: sequence effects.* Photochem Photobiol 66(2): 171-9.

[EG01] Ewens, W. J. and Grant, G. (2001). *Statistical methods in bioinformatics : an introduction.* New York, Springer.

[CB95] Karlin, S. and Burge, C. (1995). *Dinucleotide relative abundance extremes: a genomic signature.* Trends Genet 11(7): 283-90.

[KM97] Karlin, S. and Mrazek, J. (1997). *Compositional differences within and between eukaryotic genomes.* Proc Natl Acad Sci U S A 94(19): 10227-32.

[KMC97] Karlin, S., Mrazek, J. and Campbell, A. M. (1997). *Compositional biases of bacterial genomes and evolutionary implications.* J Bacteriol 179(12): 3899-913.

[LH04] Lunter, G. and Hein, J. (2004). *A nucleotide substitution model with nearest-neighbour interactions.* Bioinformatics.

[Pr92] Press, W. H., Teukolsky, S. A., Vetterling, W. T., et al. (1992). *Numerical Recipes in C, The art of scientific computing.* Cambridge, Cambridge University Press.

[RR80] Razin, A. and Riggs, A. D. (1980). *DNA methylation and gene function.* Science 210(4470): 604-10.

[Ru76] Russell, G. J., Walker, P. M., Elton, R. A., et al. (1976). *Doublet frequency analysis of fractionated vertebrate nuclear DNA.* J Mol Biol 108(1): 1-23.

[RS77] Russell, G. J. and Subak-Sharpe, J. H. (1977). *Similarity of the general designs of protochordates and invertebrates.* Nature 266(5602): 533-6.