Phylogenetic Profiling of Insertions and Deletions in Vertebrate Genomes

Sagi Snir and Lior Pachter

Department of Mathematics,
University of California, Berkeley, CA
{ssagi, lpachter}@math.berkeley.edu

Abstract. Micro-indels are small insertion or deletion events (indels) that occur during genome evolution. The study of micro-indels is important, both in order to better understand the underlying biological mechanisms, and also for improving the evolutionary models used in sequence alignment and phylogenetic analysis. The inference of micro-indels from multiple sequence alignments of related genomes poses a difficult computational problem, and is far more complicated than the related task of inferring the history of point mutations. We introduce a tree alignment based approach that is suitable for working with multiple genomes and that emphasizes the concept of *indel history*. By working with an appropriately restricted alignment model, we are able to propose an algorithm for inferring the optimal indel history of homologous sequences that is efficient for practical problems. Using data from the ENCODE project as well as related sequences from multiple primates, we are able to compare and contrast indel events in both coding and non-coding regions. The ability to work with multiple sequences allows us to refute a previous claim that indel rates are approximately fixed even when the mutation rate changes, and allows us to show that indel events are not neutral. In particular, we identify indel hotspots in the human genome.

1 Introduction

Sequence insertion and deletion events (indels) play a major role in shaping the evolution of genomes. Such events range in scale from transposable element replication within genomes, to single nucleotide events. Despite the importance of indels in modifying the function of genes and genomes [5, 24, 26], the underlying biological mechanisms are not well understood [12]. This is particularly true of small indels, also called micro-indels [25]. Analysis of micro-indels has also been limited by the availability of tractable models of indel evolution. Examples of statistical models of micro-indels include the TKF model [27], and others [17, 18, 19], however, in contrast to the large literature on evolutionary models of point mutations [11], there has been far less work on micro-indels.

The difficulty in inferring the history of insertions and deletions from a multiple sequence alignment is illustrated by a simple example. Consider a tree on three taxa (Figure 1, where the top leaf is human, the middle leaf mouse and the

A. Apostolico et al. (Eds.): RECOMB 2006, LNBI 3909, pp. 265–280, 2006.

bottom rat) and four events: two speciation events and two micro-indel events, but no point substitution event. Suppose that the primates-rodent ancestor consists of three bases. Upon the primates-rodents speciation, both ancestors keep the same three bases. Next comes the rat-mouse speciation which is followed by two parallel events: a deletion of all the three bases in the rat and a deletion of the middle base in the mouse. There is no indel event along the branch leading to the human. The true alignment of this section in the three species human, rat and mouse consists of the three human bases aligned with three gaps in the rat and a base-gap-base sequence at the mouse. In order to trace the optimal history, one may consider a site-by-site approach, however the resulting optimal sequence at the ancestral rodent is base-gap-base, yielding a history of three indel events: two deletions at sites 1 and 3 along the rat lineage and one deletion along the rodents ancestor lineage (or alternatively an insertion along the human lineage) at site 2. Obviously, this is not the true history, nor the most parsimonious one.

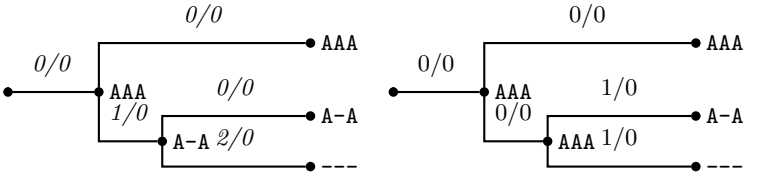


Fig. 1. An example of an alignment and two histories

The parsimony model for indel analysis has been avoided in large scale analyses, in part because the naive algorithm for reconstructing an indel history requires time that is exponential in the length of the alignment. One of our main contributions is implementing an algorithm whose running time is exponential in the number of sequences, but not in their length. This observation has already been utilized in the simplest cases. For example, as part of a broad analysis of micro-evolutionary features across the genomes of the human, mouse and rat, micro-indels and their variability was studied in [4]. It was found that there is a constant ratio between the rates of indel events and point substitutions along the mouse and rat lineages. One of the issues in such a study is the relevance of alignment quality to the results of the indel analysis, an issue which is discussed at length in [16] and which we return to when discussing indels among primates. The paper [25] restricts analysis to human and rodent coding sequences, in particular to 8148 orthologous genes among the three genomes. Only codon indel events were examined. Among the main findings was that slippage like indel events [13] are substantially more frequent than expected.

Micro-indels have also been considered in the context of reconstructing large portions of the ancestral mammalian genome [1]. Although the main goal was not the study of indels, this work was the first to deal with a non-trivial set of species and large datasets. For their purpose, a heuristic was devised in order to infer a plausible indel history and subsequently reconstruct the ancestral sequences at

gap-less positions. Although this heuristic is accurate in general, it can fail to reconstruct the true history.

In this paper we introduce the notion of an *indel history*. We assert that such a history can explain the sequence of events that occurred during the evolution of a set of species, via inference from a multiple alignment of the respective genome sequences. Our model of evolution is a restricted tree alignment model, where gap extensions receive no penalty. We argue that this specialization is biologically interesting, and computationally appealing. In particular, we develop, via a series of simplifications, an algorithm for inferring an indel history that is linear in the number of events. We also discuss the possibilities and limitations of approximation algorithms for our problem.

We applied the algorithm to coding data from the ENCODE project as well as non-coding sequences from multiple alignment of primates. Working with primates is important, as it improves the reliability of alignments [2] which are crucial for obtaining meaningful results in analysis of indels. Our findings extend the results of [4, 25] and we compare and contrast indel events in both coding and non-coding regions. The ability to work with multiple sequences allows us to refute an assumption made in [22] that indel rates are approximately fixed even when the mutation rate changes (also observed by [1]), and allows us to show that indel events are not neutral.

2 Notations and Definitions

Let us denote by $\Sigma_S = \{A, C, G, T\}$ and $\Sigma_A = \{*, -\}$. A *multiple alignment* $\mathbf{a} = a^1, \ldots, a^m$ consists of a set of m sequences with $a^i \in (\Sigma_A)^n$. We use the notation a^i_j to denote the j-th element of a^i and $[a]_j$, the *j-th column*, to denote the set a^i_j $1 \leq i \leq m$. We say that $\mathbf{a}$ has size m and length n. A multiple alignment $\mathbf{a}$ describes homology between a set of sequences $\mathbf{s} = s^1, \ldots, s^m$, $s^i \in (\Sigma_S)^{|\{j:a^i_j=*\}|}$, where each sequence element is associated with a $*$ in $\mathbf{a}$. $a^i_j = a^{i'}_j = *$ means that two elements, from sequences i and i' respectively, are 'matched' in the multiple alignment. Let $X = \{1, \ldots, m\}$ be our set of taxa and T a phylogenetic tree with leaves X. Let $\mathbf{a}$ be a multiple alignment of size m and length n An *insertion-deletion history* h (or *indel history*) consists of a labeling of vertices of T with sequences, such that each internal vertex v is labeled by a sequence $a^v \in (\Sigma_A)^n$ and each leaf i is labeled by a^i. We consider T to be a directed graph with edges directed away from the root.

Indel histories are therefore records of insertion and deletion events. Every time a $*$ switches to a $-$ there has been a deletion, and similarly every switch of a $-$ to a $*$ corresponds to an insertion. An *insertion event* corresponds to a sequence of consecutive (along one sequence) changes from $-$ to $*$, whereas a *deletion event* corresponds to consecutive (along one sequence) changes from $*$ to $-$. A history explanation, to be defined formally in the sequel, associates indel events to the given history. Let P_n denote the path of length n. Observe that we can view a history h as a function from the graph product $T \times P_n$ to Σ_A where for $v \in V(T)$ and $1 \leq j \leq n$, $h((v, j)) = a^v_j$. Let $G = T \times P_n$ and $slice_{j,G}$ (or just

$slice_j$ for short) be the graph induced by the set $\{(v,j) \in V(G)\}$. We extend the notion of parenthood of trees to G as follows: we say that $x = (v,j) \in G$ is the *parent* of $x' = (v',j') \in G$, or $x = p(x')$, if $j = j'$ and $(v \to v') \in E(T)$ (or, by the definition of $E(G)$, $(x \to x') \in E(G)$). In the sequel, we will interchangeably refer to a node either as a node in the graph, or as a combination of a tree node and an index in the path.

A leaf in G is defined analogously as in trees: a node with out-degree zero. Let $I(G)$ ($L(G)$) be the internal (leaf) nodes of G. Observe that for $j \in P_n$, $x = (v,j)$ is an internal (leaf) node in G if and only if x is an internal (leaf) node in T. Let r be the root of T and set $R = \{(r,j) \in V(G) : \ 1 \le j \le n\}$ to be the roots in G.

A convex coloring C of a graph G is a mapping of vertices of G to colors, i.e., $C : V(G) \to \{1,\ldots,k\}$ such that for each color c, the subgraph of G induced by the vertices $\{v : C(v) = c\}$ is connected [6]. We will use the notation $|C| = k$ for the number of colors in the coloring.

Given a history h, an explanation to h assigns different colors to indel events under the following rules: Two neighboring nodes in G, $x = (v,j)$ and $x' = (v',j')$ can have the same color if either $v = v' =$ the root $r \in V(T)$, or $x = p(x')$ and $h(x) = h(x')$, or $v = v'$, $j' = j - 1$, $h(x) = h(x')$ and $h(p(x')) \neq h(x')$. In addition, we require the coloring induced by the explanation to be convex on G. It is easy to see that even the naive explanation where every vertex has a different color is legal. However, we are interested in the explanation(s) with minimal number of colors. The following algorithm produces a coloring from a given history h:

1. Begin by coloring the path $r \times P_n$ monochromatically, i.e., let $C((r,j)) = 1$ for all $j \in P_n$.
2. Given a vertex $v_1 \in T$ for which all the vertices $(v_1,j), j \in P_n$ have been colored, and a child v_2 of v_1, we color the vertices $(v_2,j'), j' \in P_n$ as follows: First partition P_n into three sets $S_1 = \{j' : h(v_1,j') = h(v_2,j')\}$, $S_2 = \{j' : h(v_1,j') = * \wedge h(v_2,j') = -\}$, $S_3 = \{j' : h(v_1,j') = - \wedge h(v_2,j') = *\}$. Now set $C(v_2,j') \leftarrow C(v_1,j')$ if $j' \in S_1$. Then color each connected component of $v_2 \times S_2$ or $v_2 \times S_3$ with a unique new color (so that components get different colors from each other and from previously assigned colors). Thus, the number of new colors in C after assigning colors to $v_2 \times P_n$ is equal to the number of connected components in S_2 plus the number of connected components in S_3.

Observation 1. *The coloring obtained by the above algorithm is optimal and unique (up to the choice of colors). The number of colors corresponds to the number of indels required to explain the given history.*

By the observation above, we identify every history h with its optimal coloring, C_h. Our problem is to find the indel history h_{opt} and associated indel coloring $C_{h_{opt}}$ for which $|C_{h_{opt}}|$ is minimized.

3 Algorithm

In this section we first describe an algorithm that runs in time exponential in the number of species in the alignment and linear in the alignment length. Specifically, for m the number of species and n the length of the alignment, our algorithm runs in time $O(2^{2m-2}n)$. We then explain an improvement of the algorithm that reduces the linear factor significantly.

Let h be a history. For the purpose of the algorithm, h can be viewed as an assignment of $\{0,1\}$ to the nodes of G where 0 corresponds to a gap and 1 corresponds to an existing character state. Therefore, from now on we identify a history with its corresponding assignment. We denote by $U \subseteq V(G)$, $h|_U$ the restriction of h to the vertices of U. Recall that we index the species (the tree leaves/alignment sequences) with i and the columns of the alignment/history with j. We call $h_j = h|_{slice(j)}$ a *history slice*. We say that history slice s is *valid for* j if for every $i \in L(T)$, $s(i) = 1$ if $a^i_j = *$ and $s(i) = 0$ otherwise (that is, the slice s is consistent with the alignment at the leaves). A history h is *valid* if for every j, the history slice h_j is valid for j. Henceforth, we will restrict ourselves to valid histories and slices only. We denote by $pref(G, j)$ (or $pref(j)$ for short), the subgraph of G induced by slices $1 \dots j$.

Definition 1. *For $1 \leq j \leq n$, and a history slice s which is valid for j, let:*

$$opt(G, j, s) = \min_{h' \,:h'_j = s} |C_{h'|_{pref(j)}}|.$$

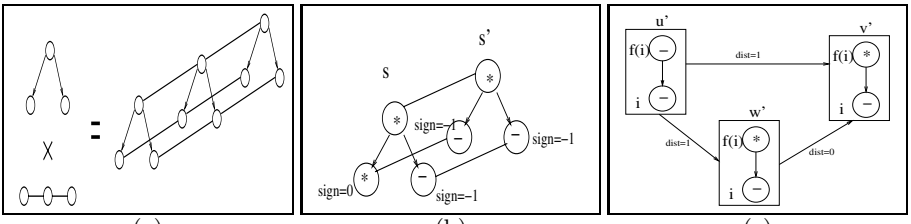
That is, $opt(G, j, s)$ is the value of an optimal history (with the least number of colors) over $pref(j)$ among all histories h' such that the j-th slice equals to s. In the sequel, we will remove G from the notation as it is clear by the context. Let $opt(j)$ be the optimal history for $pref(j)$. Since $opt(j) = \min_{s'} opt(j, s')$, the answer to the optimal indel history problem $|C_{h_{opt}}|$, is $opt(n)$. For a vertex $x \in V(G) \setminus R$ and a history h, the *sign of x under h*, $sign(h, x)$, is defined by $h(x) - h(p(x))$. In the context of slices, sign is defined for vertices of T (by omitting the index of the path).

For two history slices s and s', we have

Definition 2. $dist(s, s') = \sum_{v \in (T) \setminus r} |sign(s', v)| \delta_{sign(s,v), sign(s',v)}$.

where δ_{x_1,x_2} is the complement of the Kronecker delta (i.e. δ_{x_1,x_2} is one if $x_1 \neq x_2$ and zero otherwise). The distance between two assigned slices $dist(s, s')$ is just the sum over all vertices $v \in V(T)$, where a vertex contributes to the distance if (1) it has a different assignment than its father under s' (i.e. $sign(s', v) \neq 0$) *and* (2) it has a different sign under s' than under s (see Figure 2(b)) . Note that this distance function is *not* symmetric and therefore is not a metric. This leads us to the following observation:

Observation 2. *For $1 \leq j \leq n$, and a valid history slice for j and s, $opt(j, s)$= $\min_{s'}(opt(j-1, s') + dist(s', s))$ where $opt(0, s)$=$\sum_{v \in V(T) \setminus \{r\}} |sign(s, v)|$.*



(a) (b) (c)

Fig. 2. (a) The graph product of a cherry and P_3. (b) The distance between s and s' is computed by summing over all vertices except the root in s. If a vertex has sign $\neq 0$ and it has a different sign than its left brother, it contributes to the distance. (c) In order for i not to attain a new color under $s^{v'}$, it needs to have the same sign under $s^{v'}$ as under $s^{w'}$. This implies that i attains a new color on the edge from u' to w'.

We now define a new graph G' over the set of binary characters over T. Let $G' = (V', E', w)$ be a complete weighted directed graph where V' is the set of all binary characters (i.e. slices) over T. For an edge $e' = (u' \to v') \in E'$ such that $u' = s$ and $v' = s'$, $w(e') = dist(s, s')$. Practically, we can restrict ourselves to the sub graph of G' induced by vertices corresponding to valid slices, i.e., vertices $v' \in V'$ such that v' is valid for some j. Our problem can now be formulated slightly differently:

Problem 1. Given a tree T and an alignment $a = [a]_1 \dots [a]_n$, find a minimum weight path $P' = v_{j_1}, \dots, v_{j_n}$ in $G'(T)$ such that for every $v_{j_k} = s \in P'$, $s|_{L(T)} = [a]_k$.

Lemma 1. *w satisfies the triangle inequality.*

Proof. Consider $u', v', w' \in V'$ with corresponding history slices $s^{u'}$, $s^{v'}$ and $s^{w'}$. Recall, by the definition of *dist*, $w(u' \to v') = dist(s^{u'}, s^{v'})$ which is the number of vertices attaining new colors by moving from $s^{u'}$ to $s^{v'}$. Now, a vertex $v \in V(T)$ attains a new color upon moving from a slice s to s' when it has $sign(v) \neq 0$ under s' (i.e. different assignment than its father $p(v)$) and $sign(v)$ under s is not equal to $sign(v)$ under s' (see Figure 2(b)). Consider now the path (u', w', v') in G' (see Figure 2(c)). Let $i \in V(T)$ be a vertex such that i changes its color by moving from u' to v'. Observe that i has a different assignment than $p(i)$ (i.e. $sign(i) \neq 0$) under $s^{v'}$. In order for i to use an existing color upon moving from w' to v' on the path (u', w', v'), there must be that $sign(i)$ under $s^{w'}$ is equal to $sign(i)$ under $s^{v'}$. This implies that $sign(i)$ under $s^{w'}$ is not equal to $sign(i)$ under $s^{u'}$ *and also* $sign(i) \neq 0$ under $s^{w'}$, implying i obtains a new color upon moving from u' to w'.

Let $\mathbf{a} = ([a]_1, \dots, [a]_n)$ be an alignment. Then $\mathbf{a'}$ is a *subalignment* if it contains a subset of the columns $\{[a]_1 \dots, [a]_n\}$ in the same order as in $\mathbf{a}$. The following corollary follows from the lemma above:

Corollary 1. *Let X be a set of species, and $\mathbf{a}$ and T are an alignment and a phylogenetic tree (resp.) over X with a' a subalignment of $\mathbf{a}$. Then, $opt(\mathbf{a}, T) \geq opt(\mathbf{a'}, T)$.*

Claim. Let $\mathbf{a}^*$ be a subalignment of $\mathbf{a}$ obtained by removing every column $[a]_i$ such that $[a]_i = [a]_{i-1}$. Then $opt(\mathbf{a}, T) = opt(\mathbf{a}', T)$.

Proof. $\geq$**:** By Corollary 1.
$\leq$**:** Let H^* be the history attaining the minimal cost on $\mathbf{a}'$. It can be noted that, for every removed column, assigning the same history slice as the remaining column in $\mathbf{a}'$ under H^* (note that of every block of identical columns, exactly one remains in $\mathbf{a}'$), achieves the same cost.

Observation 2 and Claim 3 give rise to a straightforward dynamic programming algorithm which runs in time $O(2^{2m-2}n^*)$.

IndelHistory(**a**,T=(V,E))

1. remove identical adjacent columns and let n^* be the new length.
2. for every slice s valid for column 0, $opt(0, s) \leftarrow \sum_{v \in V(T) \setminus \{r\}} |sign(s, v)|$.
3. for i from 1 to n^*, $opt(i, s) \leftarrow \min_{s'}(opt(i-1, s') + dist(s', s))$.
4. return $\min_s(opt(n^*, s))$.

A more careful analysis allows us to give a better asymptotic bound.

Claim. Let $(\mathbf{a}, T)$ be an input to the problem and let $\mathbf{a}^*$ be its subalignment as in Claim 3 and with length n^*. Then $opt(\mathbf{a}, T) \geq \frac{n^*}{2}$.

Proof. Let h^* be an optimal history for (a^*, T). The proof is based on the observation that at every column (site) in h^*, at least one event is either starting or ending. We exclude the case of a whole gapped column as that does not occur in an alignment. We look at sites j and $j+1$. We divide into two cases:

1. One sequence does not change:
   Let a^i be a sequence s.t. $a^i_j \neq a^i_{j+1}$ and let $a^{i'}$ be a sequence s.t. $a^{i'}_j = a^{i'}_{j+1}$. Again we look at the cherry T_C induced by leaves i and i'. Then any history for the graph $G_C = T_C \times P_2$ with the above sequences at the leaves, must have one event.
2. The case when all sequences change is proved similarly.

Since it takes $O(nm)$ to process the alignment and by the above claim $opt(\mathbf{a}, T) \geq \frac{n^*}{2}$, we can bound the linear component in the running time by the size of the optimal solution. This implies that the time complexity of the IndelHistory algorithm is $O(mn + 2^{2m-2}|C_{h_{opt}}|)$.

4 Implementation

Although our algorithm has running time linear in the length of the alignment and the number of events, a major drawback is the exponential factor in the number of species. Our model is a special case of tree alignment (see, e.g., [23]) which has been extensively studied, and was shown to be NP-hard [28] (including a recent generalization by [9]). More recently, it has been show that there is a

generic 2-approximation and a PTAS for ancestral labeling on a tree for a set of objects whose distances satisfy the triangle inequality [29] . A more sophisticated algorithm [30] improved this result in the restricted setting of k-ary degree trees. The 2-approximation in both these works returns a *lifted alignment* in which the ancestral sequence at any node is obtained directly from the sequences at the leaves. The following observation establishes the applicability of the above result to our case:

Observation 3. *Let a_1 and a_2 be two sequences forming a cherry on an evolutionary tree. Then the ancestral sequence at the root of the cherry that minimizes the number of events along the cherry edges is either a_1 or a_2.*

In the context of alignment, the computationally expensive component of the tree alignment approximation algorithms is the pairwise alignment of all $\binom{m}{2}$ sequences. We achieve an immediate n-fold speedup of this step.

As was shown above, the ratio 2 approximation algorithm is obtained from a naive history. By this we mean the lifted history where ancestral sequences are identical to extant sequences. In order to obtain more realistic histories, we pursued two directions: an exact algorithm that uses fast operations and is suitable for up to 15 taxa, and a speedup which includes heuristics.

```
                        columns 0 to 39
                  0---------1---------2---------3---------
human             ATGCAGAGACCGGAGCCACGGACGCCCCG---ATGGA--C
cow               ACGCGGAA---------ACCGTCGGCTTGTCCGCGGAGCC
dog               ACGCGACGGTGGGTGCCGC-ACGGCCCCGG--GCGGA--T


                        columns 0 to 39
                  0---------1---------2---------3---------
internal 1 (root) ************************************--*
internal 0 ( → 1) ************************************--*
human ( → 1)      ******************************---*****--*
cow ( → 0)        *********gtgggtgcc******************GC*
dog ( → 0)        ********************c*********cc*****--*
```

Fig. 3. A toy example of input (top) and output (bottom) for the software. $*$ and - represent presence or absence of a base at a position respectively. Lowercase (uppercase) letters at a position indicate a deletion (insertion) of that base at that position, along the edge leading to that node.

The input and output to both algorithms is the same: an alignment file and a tree. The output is a complete history with numbers assigned to positions (vertices in G) indicating for each position, to which event it corresponds. Figure 3 shows an example output from the software. The upper part of the figure shows an alignment of length 40 and size 3. The lower part shows the output produced by the software. In this example there are five events.

Table 1. Comparative summary of findings in coding versus non-coding regions

Summary of Findings					
	total #sites	total #bases (non-gapped sites)	total #events	sum of events lengths	ratio of indels/ substitution
Coding Regions (from ENCODE)	156,607	931,938	661	2,343	0.013
Non-Coding Regions (CFTR)	209,381	1,438,546	2,851	8,160	0.13

In order to cope with large alignment sizes, we developed a heuristic that is linear in the number of taxa. The heuristic uses the Fitch algorithm [10] to infer the most parsimonious ancestral assignment of binary characters at each site. At nodes where the optimal assignment is not unique, the value at the same node at the previous site is assigned. The asymptotic running time of the heuristic is $O(nm)$. The algorithm was compared to the exact algorithm on two representative data sets: The vertebrates' coding regions and the primates' non-coding regions (see Section 5). We were interested only in the task of finding the best assignment (versus the complete task that includes reconstructing the solution and inferring all the events). The coding regions dataset contained six species and 156,607 sites which reduced to 1251 non-identical adjacent columns. The exponential-exact C code ran for 0.31 seconds and 661 events were inferred. The heuristic ran for 0.25 seconds and 776 events were inferred, which is 17% more than with the exact computation. Our primates non-coding dataset contained five species and 51,843 sites which reduced to 3435 non-identical adjacent columns. The exact algorithm inferred 1774 events in 0.21 seconds. The heuristic inferred 2057 events (16% more) in 0.5 seconds. In general, as the number of taxa increases, we expect the exponential factor to dominate the linear factor (reading/writing the alignment) which will be reflected in better times in favor of the heuristic computation. Notably, we were able to apply the exact algorithm to all instances analyzed in this paper.

An open question still to be addressed is whether there exists an algorithm that is polynomial in the number of species and the length of the alignment. Although the more general alignment problem is NP-hard, there is hope that such an algorithm exists since our problem is more restricted and is more structured.

5 Biological Findings

The data used was extracted from two sources: Alignments of coding regions of a set of vertebrates from the ENCODE project [7, 8] and alignments of non-coding regions of primates produced from sequence downloaded from the Program in Genomic Applications (PGA) database at Lawrence Berkeley National Laboratories [21]. Both datasets were aligned with MAVID [3], although in the case of coding region alignments, the results were re-aligned to ensure consistency of codon alignments. The latter was done by shifting every gap of size divisible by

3 to align with the human reading frame. We believe this in general resulted in a more accurate alignment.

5.1 Comparisons with Previous Studies

The excess of deletion over insertion has already been highlighted in previous studies of both coding [25] and non coding regions [26] . Our results are consistent with those studies but also reveal differences between the two types of regions. Since Taylor et al. considered only codon insertion and deletion events (i.e., indels of length divisible by 3), we filtered out all events of other lengths. We ran our software on these alignments and obtained the number of events along all branches not emanating from the root. The distribution of events obtained along the tree branches is shown in Figure 4 (right tree).

The first value we examined is the ratio between insertions and deletions along each branch of the tree. The del/ins ratio at the mouse lineage is 1.05 (versus 1.1 obtained by Taylor et. al.) and 1.50 at the rat (versus 1.7 obtained by Taylor et. al.). The second value we measured is the frequency of events along a sequence. This measure should not be confused with the rate of events along a branch in the tree. The latter indeed measures the number of events with respect to the edge length, while the frequency of events ignores this factor. In our data, there were 108,000 codons (twice the alignments length, 156,000 for both rat and mouse divided by 3) for the rat and mouse sequences and total of 73 events (sum of events for rat and mouse), yielding a frequency of one event per 1,479 codons (versus 1,736 obtained by Taylor et. al.). The agreement is striking considering that our trees contain more than twice the number of species and four-fold more branches (eight vs. two, not counting branches emanating from the root), and events could have been attributed to other branches of the tree. We now elaborate on the above argument. While in Taylor et. al. a gap in the mouse and human is automatically inferred as an insertion in the rat, in our method, based on the whole set of sequences, this scenario can be interpreted as a multi event site (see exact definition in the sequel) in which both mouse and human exhibit two different deletion events at that site.

Cooper et. al [4] found a constant ratio between the rate of indel events (insertions and deletions) measured as the number of events per site, to the rate of point substitutions per site (expressed as the length of the tree branch). This ratio, calculated only in the two mouse and rat branches and along the whole genome, was found to be 0.05. Since our non-coding data was comprised of closer species, we cannot make an exact comparison. However, the value we obtained at the rodents in coding regions was 0.0073 (obtained by summing the number of events for rat and mouse, normalized by the alignment length and divide by the length of the rat-mouse path. See values at Figure 4). Considering a ratio of 10 between coding to non-coding regions (see values at Table 1), we obtain approximately a ratio of 0.07 for non-coding regions. Taking into account the distance between human and the rodents which may lead to alignment inaccuracies, we believe the agreement is satisfactory.

5.2 Events in Coding Regions

Table 2 illustrates our finding regarding events in both non-coding and coding regions. Information regarding coding regions is the left number at every column. Both insertions and deletions decay exponentially in length. An exception to this exponential decay in the length is events of 7 codons (21 bases) that stand out in both insertions and deletions (not shown in table). We do not have an explanation for this. It can also be seen that deletions are slightly longer on average than insertions.

We also wanted to measure if, and how, the rate of indels changes along the branches of the tree. We normalized the number of events on each tree edge, by the length of the edge. This measurement enables us to estimate the correlation between the length of an edge (the expected number of substitutions on the edge) and the number of indel events accumulated on it. Another question we examined is whether the indel process is homogeneous over time, or changes along different lineages of the tree. Our coding data is composed of many genes and different

Table 2. Length distribution of indel events in coding/non-coding regions

Events in Coding/Non-Coding Regions						
	total events distribution		insertions distribution		deletions distribution	
event length	#events	total length	#events	total length	#events	total length
1	-/1895	-/1895	-/174	-/174	-/1721	-/1721
2	-/606	-/1212	-/33	-/66	-/573	-/1146
3	578/388	1734/1164	132/35	396/105	446/353	1338/1059
4	-/379	-/1516	-/20	-/80	-/359	-/1436
5	-/175	-/875	-/12	-/60	-/163	-/815
6	177/123	1062/738	48/11	288/66	129/112	774/672
7-8	-/211	-/1584	-/18	-/138	-/193	-/1446
9	55/66	495/594	11/3	99/27	44/63	396/567
10-11	-/151	-/1577	-/12	-/126	-/139	-/1451
12	56/72	672/864	7/1	84/12	49/71	588/852
13-18	34/224	543/3403	3/20	45/302	31/204	498/3101
19-30	42/175	945/4112	13/23	297/541	29/152	648/3571
total	942/4465	5451/19534	214/362	1209/1697	728/4103	4242/17837
average event length		5.786/4.37		5.649/4.68		5.826/4.34

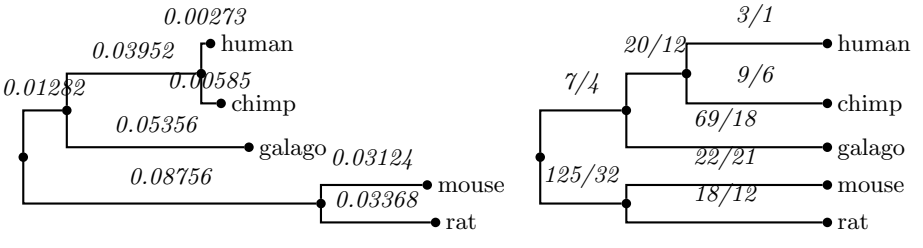


Fig. 4. Point mutation (left) and indel (right) statistics along tree edges for coding regions. The indels were computed over 156,000 sites.

sets of species. In order to obtain enough information we concatenated many genes over a common set of species. Figure 4 at the right shows the number of del/ins along each edge of the tree. We used the same data in order to obtain the edge lengths corresponding to the tree (by ML estimate according to the HKY [14] model). As the dog served as an outgroup to the rest of the species it was excluded from the figure. The tree on the left of the figure shows the edge lengths inferred for the tree (measured by the expected number of substitutions along an edge). The correlation between the edge length and the total number of events (insertions plus deletions) is notable. Specifically, for every edge e, we computed the ratio $\frac{id_e}{l_e}$ where id_e is the expected number of indel events (total number of events divided by sequence length) per site along e and l_e is the length of e (the expected number of substitution per site along it). We found that $\frac{id_e}{l_e}$ is centered around mean 0.009 (std. dev. 0.0038) with a ratio of 3.2 between the lowest value (0.005 for the ancestor of human chimp) and the highest (0.0165 for the pendant edge of the chimp). This should be contrasted to a ratio of 40 between the number of events along the pendant edge of the human (4) and 157 along the edge leading to the rodent ancestral vertex.

In [22] it was postulated that the indel process obeys the rule of molecular clock. This means that if we measure the length of the path along the tree, from any internal vertex to any of its descendants, this length will be the same. It is well known [31] that with respect to point substitutions, this hypothesis does

Table 3. Amino acid indel events

Indel Events for Amino Acids							
AA	#ins.	#del.	percent in insertions	percent in deletions	percent in population	relative insertion	relative deletion
A	45	133	10.56	9.38	7.41	1.41	1.26
C	10	14	2.34	0.98	1.73	1.34	0.56
D	10	48	2.34	3.38	4.59	0.5	0.73
E	28	117	6.57	8.25	7.16	0.91	1.15
F	5	20	1.17	1.41	3.52	0.33	0.4
G	49	129	11.50	9.1	6.56	1.73	1.38
H	15	41	3.52	2.89	2.47	1.41	1.16
I	5	23	1.17	1.62	4.05	0.28	0.39
K	19	47	4.46	3.31	5.37	0.82	0.61
L	30	143	7.04	10.09	9.89	0.7	1.02
M	5	23	1.17	1.62	2.47	0.47	0.65
N	16	55	3.75	3.88	3.24	1.14	1.19
P	39	116	9.15	8.18	6.78	1.33	1.2
Q	24	108	5.63	7.62	4.82	1.15	1.58
R	15	52	3.52	3.66	5.91	0.59	0.62
S	49	166	11.50	11.71	8.48	1.34	1.38
T	24	79	5.63	5.57	5.4	1.03	1.03
V	33	73	7.74	5.15	6.28	1.22	0.81
W	3	10	0.70	0.7	1.13	0.61	0.62
Y	2	20	0.46	1.41	2.46	0.18	0.57

not apply to the set of species we investigated here. There was acceleration in the rate of mutations in the rodents' lineage after the speciation event from the primates. This causes a substitution rate twice as much bigger in the rodents, than as in primates. Our findings refute this hypothesis. It can easily be seen that the number of events on the path from the root to the mouse is exactly 200 while to the human it is only 47. We comment here that although there are deviations in the $\frac{id_e}{l_e}$ ratio that might explain small differences, the difference here in the number of events is statistically significant.

At the amino acid level, we examined whether there was a preference for certain kinds of insertions or deletions. The composition of amino acids in insertion and deletion events is depicted in Tables 3. We inferred amino acid insertion and deletion events in both extant species (i.e. in the aligned sequences) and the ancestral nodes. An event is determined to be an insertion/deletion by the optimal explanation. It is notable that some amino acids maintain the same ratio in both processes (e.g. Arginine, Serine, Threonine) although this deviates from a neutral rate of relative value of one (e.g Arginine, Serine, Phenylalanine). Another characteristic is that most of the amino acids are either overrepresented or underrepresented in both insertions and deletions. Exceptions include Cysteine, Valine, and Glutamic acid that are over represented in one process but underrepresented in the other.

5.3 Events in Non-coding Regions

Our non-coding data was taken from homologous sequences of primates surrounding various genes (see [21]). Here the emphasis was to examine the deleted and inserted sequences and their properties. Values are shown in Table 2 (right number of every column). There are 4465 events with total length of 19534 bases, which yields an average event length of 4.37 bases per event. Events of a single base comprise 42.4 percent of the total number of events and of length two, 13.5 percent. Of the total number of events, there are 362 insertion events with total length of 1697, yielding an average insertion size of 4.68 bases. In turn, there are 4103 deletion events with total size of 17837, yielding average deletion size of 4.34 bases.

Table 4 shows the base composition of indels in non-coding regions. We used the same method here for the inference of the content of the indel events as we did for coding regions, except for the fact that we considered indels of all length. We found that the percentage of Gs and Cs in indel events was even lower than the population GC content. In insertion events C is substantially underrepresented (0.81 of its background frequency) while T is similarly overrepresented (1.13). In deletion events, both A and T are similarly overrepresented (around 1.05 of their ground frequency) while C and G are similarly underrepresented (0.92). C and T exhibit the largest variation between insertion and deletions.

Similarly to coding regions, we wanted to measure the correlation between rate of indel events to the rate of point substitutions along the tree branches. Figure 5 depicts our findings in the CFTR region (ENCODE region number 1). The right tree depicts the distribution along the edges. The edge lengths of

Table 4. Distribution of bases in insertions and deletion events in non-coding regions

Bases Distribution in Non-Coding Regions					
base	% in population	% in insertions	% in ins relative to % in population	% in deletions	% in ins del relative to % in population
A	28.3	30.6	1.08	30.3	1.06
T	29.5	33.3	1.13	30.8	1.04
C	20.8	17.0	0.81	19.1	0.92
G	21.2	18.9	0.89	19.6	0.92

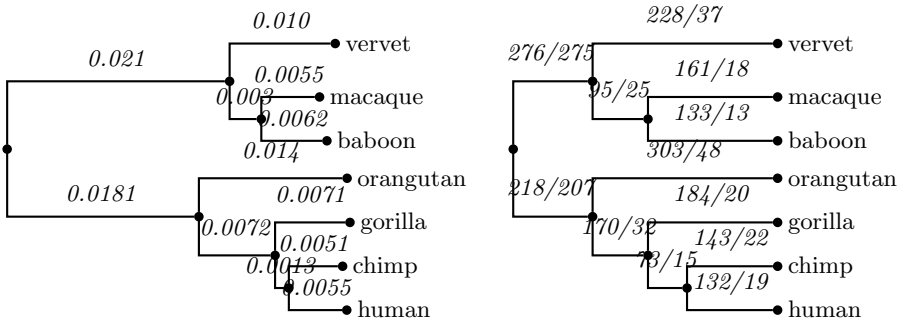


Fig. 5. Point mutation (left) and indel (right) statistics along tree edges for the CFTR region. The indels were computed over 209,000 sites.

the tree in the left correspond to the point substitution probabilities. Here the value $\frac{id_e}{l_e}$ (see definition in the coding region section) is centered around a mean of 0.146 (std. dev. 0.061) with a single big exception for the ancestral edge of human and chimp which is double that value.

5.4 Indel Hotspots

Multi event sites (MES) are sites where an indel event has occurred on more than one branch of the tree. Indel events at MES sites are called parallel events.

Our findings show that in both datasets, the frequency of parallel events was more than two fold above its expected value. Specifically, for coding regions, the number of sites containing a single event was 9,553 yielding an expected value of 0.0295 (recall the total number of sites was 323,673) and a probability of 0.000871 of finding a parallel event at a site. The actual number of parallel events was 1093, yielding a frequency of 0.00337 parallel events per site, 3.876 times its expected value. For non-coding regions, we found 30,616 sites containing an event, yielding a frequency of 0.0714 sites containing events, and a probability of 0.0051 for a parallel event at a site. The actual frequency of parallel events was 0.0111, which is 2.178 times its expected value.

These findings are consistent with the findings in [25] about the effect of slippage at indel events. [25] found that the frequency of indel events is about

50% higher than expected in proximity to small regions where the same amino acids is duplicated multiple times. As we have shown, such indel "hotspots", are also evident in non-coding sites. Although some indel hotspots may be due to alignment artifacts as suggested in [16], we believe that our results confirm that indel hotspots exist.

Acknowledgments

SS and LP were partially funded by a grant from the NIH (R01: HG2362-3). We thank Yifat Felder, Benny Chor, and the anonymous referees for comments that improved the manuscript. We also thank Colin Dewey for his generous help.

References

1. Blanchette, M., Green, E. D., Miller, W., Haussler, D. (2004). Reconstructing large regions of an ancestral mammalian genome in silico. Genome Res. 14, p 2412-2423.
2. Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., Pachter, L., Rubin, E.M. (2003). Phylogenetic shadowing of primate sequences to find functional regions of the human genome, Science, Volume 299, Number 5611 (2003), p 1391–1394.
3. Bray, N. and Pachter, L.. (2004) MAVID: Constrained ancestral alignment of multiple sequences Genome Res. 14, p 693-699.
4. Cooper, G.M., Brudno, M., Stone, E.A., Dubchak, I., Batzoglou, S., and Sidow, A. (2004). Characterization of evolutionary rates and constraints in three mammalian genomes. Genome Res. 14, p 539-548.
5. Chuzhanova N.A., Anassis E.J., Ball E.V., Krawczak M., Cooper D.N. (2003), Meta-analysis of indels causing human genetic disease: mechanisms of mutagenesis and the role of local DNA sequence complexity. Human Mutation 21(1), p 28-44.
6. Dress, A., Steel, M.A. (1992). Convex tree realizations of partitions. Applied Mathematics Letters 5(3), p 3–6,
7. The ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia of DNA Elements) Project. Science, 306(5696), p 636–640.
8. The Berkeley ENCODE Website. http://bio.math.berkeley.edu/encode/
9. Elias, I. (2003). Settling the Intractability of Multiple Alignment. Int. Symp. on Algorithms and Computation (ISAAC) p 352-363.
10. Fitch, W.M. (1981). A non-sequential method for constructing trees and hierarchical classifications. J. Mol. Evol. 18(1), p 30–37.
11. Felsenstein, J. (2004). Inferring Phylogenies. Sinauer Associates Inc., Mass.
12. Frazer K.A., Chen X., Hinds D.A., Pant P.V., Patil N., Cox D.R. (2003). Genomic DNA insertions and deletions occur frequently between humans and nonhuman primates. Genome Res. 13(3), p 341-6.
13. Hancock J.M., Vogler A.P. (2000). How slippage-derived sequences are incorporated into rRNA variable-region secondary structure: Implications for phylogeny reconstruction. Mol. Phylogenet. Evol. 14, p 366-374.
14. Hasegawa, M., Kishino H., and Yano, T. (1985). Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22, p 160–174

15. Lai, Y. and Sun, F. (2003). The relationship between microsatellite slippage mutation rate and the number of repeat units. Mol. Biol. Evol. 20, p 2123-2131
16. Löytynoja, A. and Goldman, N. (2005). An algorithm for progressive multiple alignment of sequences with insertions. Proc. Natl. Acad. Sci. 102,p 10557-10562.
17. McGuire, G., Denham, M.C., and Balding, D.J. (2001). Models of sequence evolution for DNA sequences containing gaps. Mol. Biol. Evol. 18, p 481-490.
18. Mitchison, G. J. (1999). A probabilistic treatment of phylogeny and sequence alignment. J. Mol. Evol. 49, p 11-22.
19. Mitchison, G. J., and Durbin R. M. (1995). Tree-based maximal likelihood substitution matrices and hidden Markov models. J. Mol. Evol. 41, p 1139-1151.
20. Petrov, D.A., Sangster, T.A., Johnston, J.S., Hartl, D.L., and Shaw, K.L. (2000). Evidence for DNA loss as a determinant of genome size. Science 287, p 1060-1062.
21. Berkeley PGA. http://pga.lbl.gov/
22. Saitou N. and Ueda S. (1994). Evolutionary rates of insertion and deletion in noncoding nucleotide sequences of primates. Mol. Biol. Evol. 11(3), p 504-12.
23. Sankoff, D., and Cedergren, R. (1983). Simultaneous comparisons of three or more sequences related by a tree, in D. Sankoff and J. Kruskal (eds), Time Warp, String Edits, and Macromolecules: the Theory and Practice of Sequence Comparison, p 253-264, Addison Wesley, Reading Mass.
24. Soding J. and Lupas A.N. (2003). More than the sum of their parts: on the evolution of proteins from peptides, Bioessays, Sep 25(9), p 837-46.
25. Taylor, M.S., Ponting, C.P., Copley, R.R. (2004). Occurrence and consequences of coding sequence insertions and deletions in mammalian genomes Genome Res. 14, p 555–566.
26. Thomas, J.W., Touchman, J.W., Blakesley, R.W., Bouffard, G.G., Beckstrom-Sternberg, S.M., Margulies, E.H., Blanchette, M., Siepel, A.C., Thomas, P.J., McDowell, J.C., et al. (2003). Comparative analyses of multi-species sequences from targeted genomic regions. Nature 424, p 788-793.
27. Thorne, J. L., Kishino H, and Felsenstein J. (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. J. Mol. Evol. 33, p 114-124.
28. Wang, L., and Jiang, T. (1994). On the complexity of multiple sequence alignment. Journal of Computational Biology 1(4), p 337-348.
29. Wang, L., Jiang, T., and Lawler, E.L. (1996). Approximation algorithms for tree alignment with a given phylogeny. Algorithmica 16(3), p 302-315.
30. Wang, L., and Gusfield, D. (1997). Improved approximation algorithms for tree alignment. J. Algorithms 25(2), p 255-273.
31. Wu, C. and Li W.H. (1985). evidence for higher rates of nucleotide substitution in rodents than in man. Proc. Natl. Acad. Sci. 82, p 1741-1745.