Inferring indel parameters using a simulation-based approach

Eli Levy Karin¹, Avigayel Rabin¹, Haim Ashkenazy¹, Dafna Shked⁵, Oren Avram¹,², Reed A. Cartwright³,⁴ and Tal Pupko§¹

¹ Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel-Aviv University, Tel-Aviv 69978, Israel
² The Blavatnik School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel
³ The Biodesign Institute, Arizona State University, Tempe, AZ USA 85287-5301
⁴ School of Life Sciences, Arizona State University, Tempe, AZ, USA 85287-5301

*Equal contribution
§Corresponding author

Tal Pupko, Tel: 972-3-6407693; Fax: 972-3-6422046
E-mail: talp@post.tau.ac.il

Running title: Inferring indel parameters

Keywords: Simulations; Phylogeny; Indels; Alignments; Mahalanobis distance;
Abstract

In this study we present a novel methodology to infer indel parameters from multiple sequence alignments (MSAs) based on simulations. Our algorithm searches for the set of evolutionary parameters describing indel dynamics which best fits a given input MSA. In each step of the search, we use parametric bootstraps and the Mahalanobis distance to estimate how well a proposed set of parameters fits input data. Using simulations, we demonstrate that our methodology can accurately infer the indel parameters for a large variety of plausible settings. Moreover, using our methodology, we show that indel parameters substantially vary between three genomic datasets: mammals, bacteria, and retroviruses. Finally, we demonstrate how our methodology can be used to simulate MSAs based on indel parameters inferred from real datasets.
Introduction

A large body of research is dedicated to understanding how the evolutionary process varies within groups of orthologues, among sites within a gene, between populations and among diverged species. Evolutionary models aiming to describe these dynamics must account for base pair substitutions as well as insertion and deletion (indel) events.

Great progress has been made in developing rich and accurate substitution models that account for many features of the evolutionary process. Such features include accounting for differences between rates of transitions and transversions as implemented, for example, in the HKY85 model (Hasegawa, Kishino, Yano 1985), accounting for among site rate variation by using a Gamma distribution (reviewed in Yang 1996; Pupko, Mayrose 2010), considering heterotachy (e.g., Whelan, Blackburne, Spencer 2011), lifting the assumption of stationarity (Barry, Hartigan 1987; Lockhart et al. 1994), and accounting for dependencies among sites (Yang 1995; Robinson et al. 2003; Siepel, Haussler 2004; Rodrigue et al. 2005; Stern, Pupko 2006; Suzuki, Gojobori, Kumar 2009; Berard, Gueguen 2012).

Following the pioneering TKF models (Thorne, Kishino, Felsenstein 1991; Thorne, Kishino, Felsenstein 1992), several models to describe indel dynamics have been proposed: from statistical alignment algorithms (Lunter et al. 2003), through the development of the long indel model (Miklos, Lunter, Holmes 2004), to the use of the Poisson distribution to describe a fixed instantaneous rate of indels (Cartwright 2005). The length distribution of indels has been studied in various biological datasets and was
proposed to follow a Zipf (power law) distribution (Benner, Cohen, Gonnet 1993; Gu, Li 1995; Zhang, Gerstein 2003; Chang, Benner 2004; Zhang et al. 2010). This distribution describes an inverse relation between the length of an indel (k) and its probability:

\[ \text{Pr}(k) = \frac{k^{-a}}{\zeta(a)}, \]

where \( a > 1 \) is the slope parameter of the distribution, and \( \zeta(a) = \sum_{n=1}^{\infty} n^{-a} \) is the Riemann zeta function.

When utilizing substitution models, free parameters are usually inferred from data. For example, many applications rely on the inference of the parameter of the gamma distribution used to model among-sites rate variation (Buckley, Simon, Chambers 2001; Susko et al. 2003; Pond, Frost 2005; Abhiman, Daub, Sonnhammer 2006; Rubinstein et al. 2011). Another example is the evaluation of the parameter which indicates the type and intensity of the selection regime acting on a protein coding gene (e.g., Goldman, Yang 1994; Yang et al. 2000; Bielawski 2013).

In contrast to the common practice of inferring substitution related model parameters, not much work has been devoted to inferring parameters describing indel dynamics, such as the ratio between indel rates and substitution rates and the distribution of indel size. Two notable efforts to infer indel parameters include the lambda.pl script implemented as part of the Dawg package (Cartwright 2005) and an expectation maximization algorithm to infer those parameters from HMMs of a pair of sequences (Cartwright 2009). A possible reason for the paucity of analyses for which indel parameters are inferred is that indel dynamics parameters are substantially more challenging to estimate compared to substitution parameters (Cartwright 2005; Fletcher,
Yang 2009). Factors contributing to this challenge include the dependency among sites introduced by indel events and the existence of overlapping indels.

Since the indel dynamics parameters are an essential precursor to many phylogenetic procedures, we were motivated to find a way to infer these parameters from input datasets. In this study, we present a novel algorithm to infer the indel to substitution rate ratio, the parameter controlling the distribution of indel length and the root length from an input MSA and tree. Our proposed method is general and could be applied to a wide variety of theoretical models. We demonstrate our method by analyzing the indel dynamics in three datasets—from mammals, from the COG database, and from an HIV-1 dataset—demonstrating the variability of indel dynamics among protein MSAs. We further demonstrate how combining our algorithm with sequence simulators leads to simulated sequences that mimic real sequence datasets in terms of indel prevalence and length. We provide the SPARTA (Searching indel PARameters Trained from Alignment) software implementing our algorithm.

Results

The algorithm

Our algorithm takes a given MSA as input and infers three relevant parameters regarding indel dynamics: (1) IR, the indel-to-substitution rate ratio; (2) the slope parameter ‘a’ of the power-law distribution, which controls the distribution of the lengths of indels; and (3) RL, the length of the sequence at the root of the phylogeny (for more details, see Materials & Methods section). Our proposed methodology uses a simulation
based approach to search over the space of parameters for the ones that best fit the input MSA. In each search step, a specific set of input parameters is estimated. This set is refined (using standard hill-climbing heuristics) until an optimal set of parameters is inferred. The fit of each set of parameters to the input MSA is computed as a measure of the distance between the input MSA and a set of MSAs produced by simulating sequences under this set of parameters. In order to calculate this distance, we compute a vector of attributes for the input MSA as well as for each of the simulated MSAs produced under the set of parameters of the current step. These attributes are summary statistics computed from each MSA (e.g., the average gap block length and the number of gap blocks, see Materials and Methods section). We next compute the Mahalanobis distance (Mahalanobis 1936) between the attribute vector of the input MSA and the distribution of attribute vectors computed for each of the $N$ simulated MSAs in the current step:

$$D(x) = \sqrt{(x - \mu)^T S^{-1} (x - \mu)}$$

where $x$ is the vector of attributes computed for the input MSA, $\mu$ is the vector of attribute means for the set of simulated MSAs, and $S$ is the covariance matrix for the set of simulated MSAs. Our working hypothesis is that this distance is a good estimate of the difference between the current search step parameters and the parameters underlying the input MSA, i.e., a large $D(x)$ value indicates that the current set of parameters is unlikely to generate MSAs with the same characteristics as the input MSA. Our parameter search procedure, SPARTA, is described schematically in Figure 1. An example of the dependency of the Mahalanobis distance surface for a grid of
parameter combination is presented in Supplementary material, Figure S1, which shows that indeed, the Mahalanobis distance increases as the parameter values are further away from the true set of parameters.

**Tuning the methodology and run time analysis**

We first set out to fine tune our methodology, which depends on the number $N$ of simulated MSAs. To this end, we simulated a set of MSAs with known parameters under a specific tree topology and branch lengths and tested the ability of our methodology to accurately infer parameters. Specifically, the tree topology and branch lengths were chosen to reflect a biological dataset (see Materials and Methods, the Azurin dataset), and different configuration sets were used to model indel dynamics. Using the “basic configuration set” (see Materials and Methods section), we simulated 50 MSAs using INDELible (Fletcher, Yang 2009). As we show below, this parameter configuration is well within the range of biologically plausible parameters. SPARTA then estimated the indel parameters of these 50 simulations. Figure 2A summarizes the dependence of the Mahalanobis distance on $N$ (and provides information regarding run times), while Figure 2B shows the distance between each “true” parameter and the inferred one, as a function of $N$. As expected, the accuracy of the parameter-search procedure increases with $N$. We chose $N = 100$ for further analyses as this value offers a good compromise between accuracy and computation time. For $N = 100$, the average inferred parameters and one standard error were indel rate = 0.023±0.006, ‘a’
parameter = 1.324±0.116, root length = 341.94±24.764. These values well fit the real parameters: 0.02, 1.3, 350.

Accuracy evaluation

We next evaluated the performance of our methodology as a function of the parameters used to simulate the “true” MSA. We aimed to determine the accuracy as a function of the above presented three parameters: low versus high number of indel events, shorter versus longer indels and different root lengths. To this end, we tested the “basic parameter configuration” as well as six alternative parameter configurations, each of which differs from the basic parameter set by the value of one parameter (see Materials and Methods). These configurations were chosen to represent a wide range of evolutionary scenarios. The results of these analyses are presented in Table 1. As can be seen, in all parameter configurations, the real parameter values fall within one standard error from the inferred parameter value. Thus, the algorithm is able to reconstruct a broad spectrum of parameter values. The results suggest an increased accuracy for high values of indel rate which may be explained by the presence of more indel events from which reliable estimates may be obtained.

Comparison to Dawg parameter inference

A previous effort to infer indel parameters is implemented as a Perl script lambda.pl which is part of the Dawg package (Cartwright 2005). We compared the performance of
parameter estimation by SPARTA with that by the Dawg package using simulations. The accuracy performance on the basic parameter configuration is shown in Figure 3. As can be seen, for this set of parameters, lambda.pl overestimates the ‘a’ parameter (i.e., shorter indels) and is thus less accurate than SPARTA. The difference in accuracy between the method for this parameter is statistically significant (comparing the squared errors between the true parameter values and the inferred ones using Wilcoxon paired test; p-value < 8.44e-08). No statistical significant difference between the two methods was found for the estimation of the indel rate parameter. SPARTA was slightly less accurate for the inference of the root length parameter (p-value < 2.7e-03). We repeated this comparison for all six alternative parameter sets (see Materials and Methods). The results (Supplementary material, Figure S2) suggest that lambda.pl tends to overestimate the ‘a’ parameter for many of the examined parameter sets. Furthermore, we counted the cases in which the real parameter values fell within a single standard error from the average inferred parameters. While for SPARTA this was the case for all sets of parameters examined, Dawg managed to correctly infer only a single complete set. These results suggest that SPARTA is a valuable alternative to Dawg for accurate inference of indel parameters.

**Inference under different alignment programs**

In all of the above described analyses, the attributes used for the inference of indel dynamics parameters were extracted from the “true” MSA. In practice, however, the true MSA is not available and an MSA has to be reconstructed from the sequence data by
an alignment program. Our next goal was to study the impact of commonly used alignment programs on the performance of SPARTA. To this end, we examined ClustalW V1.8 (Thompson, Higgins, Gibson 1994; Higgins, Thompson, Gibson 1996), MAFFT V7 (Katoh, Standley 2013) and PRANK V140603 (Loytynoja, Goldman 2008; Loytynoja 2014). We simulated sequence datasets using INDELible under the “basic parameter configuration”. The simulated unaligned sequences were given to each of the alignment programs to compute MSAs. Those MSAs were next used as input MSAs for the algorithm and the accuracy of the algorithm was evaluated. As can be seen in Figure 4, for all three alignment programs considered, the inference of the parameter and the root length were relatively accurate despite the fact that the MSA is inferred rather than known. Regarding the estimation of the indel rate, both MAFFT and PRANK allow for accurate estimation of this parameter. However, this parameter is substantially underestimated when the MSA is reconstructed using ClustalW. This is in accordance with previous reports suggesting that ClustalW tends to over-align sequences (Loytynoja, Goldman 2008; Privman, Penn, Pupko 2012). We repeated this analysis for all six alternative parameter configurations (see Materials and Methods). The results (Supplementary material, Figure S3) further support our observation for the basic parameter set, i.e. the indel rate parameter in ClustalW MSAs tends to be underestimated. Moreover, following alignment by all MSA programs, the indel rate parameter was underestimated under the high indel rate configuration, with alignments by PRANK and MAFFT yielding closer estimates to the true value compared to ClustalW.
Simulating data with SPARTA parameter estimates

Sequence simulators are used for a wide variety of phylogenetic analyses (Fletcher, Yang 2010; Bay, Bielawski 2011; Gaston, Susko, Roger 2011; Izquierdo-Carrasco, Smith, Stamatakis 2011; Wu, Susko 2011; Blackburne, Whelan 2012; Jordan, Goldman 2012; Kuck et al. 2012; Loytynoja, Vilella, Goldman 2012; Privman, Penn, Pupko 2012; Thi Nguyen, Gesell, von Haeseler 2012; Bay, Bielawski 2013; Wang, Susko, Roger 2013; Redelings 2014; Spielman, Dawson, Wilke 2014). For each sequence simulation, specific models are assumed, and investigators must provide the model parameters. Given a specific indel model used by the simulator, the task of selecting its parameters so that the simulator could produce MSAs similar in their indel characteristics to the studied MSA is not a trivial one. Notably, the selection of indel parameters for a given MSA motivated the inclusion of lambda.pl as part of the Dawg simulation package. To demonstrate the utility of SPARTA to infer such parameters, we aimed to simulate MSAs that resemble the indel characteristics of the Azurin protein MSA (Azurin dataset, see Materials and Methods) shown in Figure 5A. A typical MSA simulated based on the parameters inferred by our methodology is presented in Figure 5B. As can be seen, the real MSA and the simulated MSA are similar with respect to their total length and the number and size of indels. In contrast, a typical MSA simulated with the default INDELible parameters (Figure 5C) is characterized by shorter indels and a much longer alignment length compared to the true MSA. In both simulations, the Azurin tree topology and branch lengths were used and the same substitution model was assumed.

Indel dynamics in biological datasets
Having established that our method can accurately infer indel dynamics parameters, we next applied our method to study real biological datasets. We examined 498 mammalian MSAs obtained from the OrthoMam V8 database (Douzery et al. 2014) and 100 MSAs obtained from the COG database (Tatusov et al. 2003). We ran the search algorithm on each such MSA to infer its parameters (Figure 6). Our results show that indel dynamics differ between mammals and the COG MSAs that include a relatively diverse set of (mostly microbial) organisms. COG MSAs are characterized by a narrow distribution of short root lengths. Furthermore, the indel rate is higher in COG MSAs compared to mammalian MSAs, indicating that indel dynamics vary among different clades. Notably, when analyzing such real biological datasets, additional sources of bias regarding MSA inaccuracies exist, e.g., the inclusion of paralogs instead of orthologs, the inclusion of only partial sequences, or biases due to filtering of specific sequences and positions. Nevertheless, the above results demonstrate the utility of our method to estimate indel parameter distributions for various phylogenetic groups.

In addition, we inferred the indel dynamics parameters for nine HIV-1 coding genes. As can be seen in table 2, there is at least an order of magnitude variation in the inferred indel rate parameter, ranging between 0.0003 and 0.0195 indels to substitutions. These values are lower than those inferred for the OrthoMam and COG datasets, suggesting that indels are less common compared to substitutions for HIV-1 than for mammals and bacteria. Furthermore, experimental data suggest that indel mutations comprise of 3-6% of the total cases of mutations (Abram et al. 2010), which is approximately the values inferred for seven out of the nine HIV-1 genes we analyzed. For two HIV-1 genes, \textit{vif}
and vpr, a much lower indel rate parameter value was inferred (Table 2). This result is in line with the low number of indels compared to substitutions observed in the MSAs of these genes (the alignment of *vif* is shown in Figure 7). The low indel rate ratio of these genes compared to the other HIV-1 protein coding genes suggests these two genes are either subject to a different mutation regime or to a stronger purifying selection against the introduction of indels.

**Discussion**

In this study we presented a novel method to infer indel dynamics parameters from an input MSA by extracting summary statistics. Specifically, we demonstrated the ability of the algorithm to recover three parameters. However, our method is general and is not limited to a specific theoretical model of indels and could be, with minor adjustments, applied to study the parameters of other theoretical models describing indel dynamics. As a case in point, it can be used to study other indel size distributions, such as the Lavalette distribution (Popescu 2003; Fletcher, Yang 2009).

The last years have seen a steady progress towards advanced Bayesian approaches aimed at reconstructing trees and MSAs simultaneously (Thorne, Kishino, Felsenstein 1992; Lunter et al. 2005; Redelings, Suchard 2005; Bouchard-Cote, Jordan 2013; Herman et al. 2014). One of the strengths of these approaches is that they integrate over uncertainty in model parameters, including those relevant to indel dynamics. Thus, such methodologies can, in principle, provide posterior estimates for indel dynamics.
parameters. Unfortunately, such methods are generally computationally intensive and hence these approaches cannot be easily applied to datasets containing hundreds or thousands of taxa.

Similar to these Bayesian methodologies, the SMUVE approach (Cartwright 2009; software available at https://github.com/reedacartwright/emdel) presents a likelihood-based model to infer indel dynamics parameters, using pairwise HMMs. While this approach is clearly more powerful than previously developed methods (discussed in Cartwright 2009), it is limited to the inference from pairwise alignments only. Furthermore, the methodology infers indel parameters by aligning the two sequences, while our model takes the MSA as input.

In this study we compared our approach to a methodology for the inference of indel parameters implemented in lambda.pl, part of the Dawg package. Lambda.pl first estimates the number of unique gap characters, and from which, assuming a Poisson distribution, it infers the ratio between indel and substitution events. Further, from the estimated size frequencies of the unique gaps, it infers the parameters of the indel size distribution. The inaccuracies we observed when testing the performance of lambda.pl may stem from errors in counting unique indels (e.g., due to overlapping indel events, or parallel indel formation). Notably, SPARTA as opposed to lambda.pl, does not require a rooted tree, which is often unavailable (e.g., when gene trees are analyzed).
Furthermore, our proposed methodology opens the way to account for indels in simulation studies by obtaining indel dynamics parameters which were inferred from biological datasets and subsequently generating sequences that evolved under these parameters. Our methodology can further be used in tests based on parametric bootstrap. While previously developed parametric bootstrap methodologies have simulated sequences without indels (Bull et al. 1993; Swoford et al. 1996) it is now possible to simulate alignments with indel parameters that have been estimated from the real MSA. Parametric bootstrapping with indels may be very important in cases where alignment uncertainty may affect downstream analyses, e.g., when inferring positive selection (Jordan, Goldman 2012; Privman, Penn, Pupko 2012; Blackburne, Whelan 2013; Redelings 2014; Spielman, Dawson, Wilke 2014) and testing if two or more trees are equally supported by the data (Levy Karin, Susko, Pupko 2014). Since in many cases it is unknown if indels in the alignment affect downstream analyses, we suggest that whenever a parametric bootstrap approach is utilized, indels should be accounted for by simulating with indel parameters that are estimated from real datasets.

In addition to utilizing SPARTA together with a sequence simulator for simulating datasets that resemble an input MSA (as exemplified in Fig. 5), our method can also be utilized to detect indel dynamics parameters for specific genes or lineages. We have exemplified these possible uses by comparing indel rates among mammalian, bacterial and viral sequences (Fig. 6) and among HIV-1 genes (Fig. 7). Differences in indel dynamics may reflect changes in either the mutation process or in the selection regime.
For example, a higher indel rate in a given gene may stem from relaxation in purifying selection or, in rare cases, from positive selection.

We applied SPARTA to a large number of biological datasets and tested whether the real summary statistics were consistent with the distribution of summary statistics generated by simulations using the inferred indel parameters. A large fraction of datasets were significantly different ($\alpha = 5\%$) than their associated simulations (394 out of 498 Orthomam datasets, 72 out of 100 COG datasets, and 3 out of 9 HIV-1 MSAs). This suggests that further improvements in indel modeling are needed. For example, currently it is assumed that the indel parameters are shared across all positions, while it is plausible that indels are more likely in some regions than others. Similarly, indel dynamics may vary across tree lineages. These aspects await further research.

It should be noted that the methodology presented here infers indel dynamics from input MSAs that were themselves generated by MSA programs that assume specific indel parameters and are subjected to various biases (Lunter et al. 2008). One could claim that such an approach may be biased towards inferring the parameters used to reconstruct the input MSAs. Furthermore, it is well established that the “optimal” MSA obtained by each alignment program reflects only one possible path out of many equally likely paths and many more suboptimal solutions (e.g., Landan, Graur 2007). One possible improvement of the SPARTA methodology would be to account for MSA uncertainty by averaging the parameter estimations over a large sample of plausible
MSAs. Such alternative MSAs can be obtained using the GUIDANCE methodology (Penn et al. 2010; Sela et al. 2015) or statistical MSA methodologies (Hein et al. 2000; Herman et al. 2015). Although it would be computationally expensive, biases introduced by MSA programs can be incorporated into SPARTA by realigning simulated MSAs using the same aligner as the original dataset and comparing the results to the input MSA.

Our methodology currently estimates the best set of indel dynamics parameters that match a specific input MSA. In the future, the method can be extended to compute not only optimal parameters, but rather, a distribution of plausible parameters for each input alignment. This can be achieved using methodologies such as Approximate Bayesian Computation (Marjoram et al. 2003), which are already extensively used in population genetics studies (Blum, Jakobsson 2011; Shafer et al. 2015).
Materials and Methods

Indel dynamics parameters

In this study we demonstrate our methodology for the inference of three indel parameters. These parameters are used by the sequence simulation program to generate the evolutionary process along the tree. The first parameter is the indel-to-substitution rate ratio (IR), which controls the proportion of events in the simulation in which an indel is created. The second parameter is the 'a' parameter of the power law distribution, which controls the distribution of indel length. This distribution describes an inverse relation between the length of an indel and its probability. The third parameter is the length of the sequence at the root of the tree (RL). The simulation is of an evolutionary process along a tree, and this sets the length of the ancestral sequence which is mutated along the branches of the tree. (These parameters are described in: Fletcher, Yang 2009).

Attributes and attribute computation

1. **MSA length**: the number of columns in the alignment.
2. **Total number of gap blocks in the MSA**: gap blocks are one or more consecutive gap characters. This is an estimate of the number of indels per sequence, over the whole MSA.
3. **Average gap block length**: the total number of gap characters divided by the total number of gap blocks. This is a proxy of average indel length.
4. **Minimal length of sequence in the input MSA**
Confidence measure for SPARTA’s inferred parameters

The SPARTA methodology can be used to examine the fit between summary statistics of the input MSA and those statistics computed from the simulated MSAs under the inferred indel parameters. This is achieved by first computing the Mahalanobis distance between the vector of summary statistics computed from the real data to the multivariate distribution of summary statistics computed from the simulated MSAs. Next, this distance is translated to a p-value according to the following formula (Clark, Dunn, Smith 1993):

\[
p = 1 - \text{CDF.CHISQ}(D^2, v - 1)
\]

where \(D\) is the Mahalanobis distance, \(v\) is the number of summary statistics, and \(\text{CDF.CHISQ}\) is the chi-squared cumulative distribution function with \(v-1\) degree of freedoms.

A significant p-value (e.g., smaller than 0.05) means that the vector of summary statistics computed from the real MSA is unlikely to originate from the same multivariate distribution yielding the set of simulated MSAs, suggesting the assumed indel model does not capture some indel aspects reflecting the evolution of the real data analyzed.

Parameter configurations

In this study we examined the inference of the slope of the power-law distribution of indel lengths (a), the indel-to-substitution rate ratio (IR), and the ancestral sequence length (RL). The following parameter configurations were used in this study:
(1) “Basic configuration”: $a=1.3$, $IR=0.02$, $RL=350$

(2) Alternative configuration 1: $a=1.3$, $IR=0.02$, $RL=100$

(3) Alternative configuration 2: $a=1.3$, $IR=0.02$, $RL=500$

(4) Alternative configuration 3: $a=1.1$, $IR=0.02$, $RL=350$

(5) Alternative configuration 4: $a=1.7$, $IR=0.02$, $RL=350$

(6) Alternative configuration 5: $a=1.3$, $IR=0.01$, $RL=350$

(7) Alternative configuration 6: $a=1.3$, $IR=0.1$, $RL=350$

These configurations were simulated using INDELible. All other INDELible parameters were set to: ‘NUCLEOTIDE 2’ model, substitution model: ‘HKY 2.5’, maximum indel length: 50. For all simulations, the Azurin tree described below was used.

**Datasets**

**Azurin dataset.** The Azurin protein MSA was downloaded from the HOMESTRAD database (Mizuguchi et al. 1998). This set includes 29 sequences, the MSA length is 215 amino acids. The maximum likelihood phylogenetic tree for this dataset was reconstructed using PhyML (Guindon et al. 2010; Criscuolo 2011) with the following parameters: model of amino acids substitution = WAG, discrete gamma model, number of categories = 4, tree topology search = best of NNI and SPR, and optimizing over all other parameters (‘tlr’, proportion of invariant sites estimated from the data). The tree is given in the supplementary information.
**OrthoMam dataset.** A mammalian collection of orthologous genes was downloaded from the OrthoMam database (Douzery et al. 2014). This collection included 498 MSAs which had orthologs across all 40 mammalian sequences. The gene trees provided by OrthoMam were used in all our analyses.

**COG dataset.** A collection of orthologous genes was downloaded from the COG database (Tatusov et al. 2003). This collection included 100 MSAs. Each of these MSAs contained 40-50 sequences of genes. The maximum likelihood gene tree for each set of orthologs was reconstructed using PhyML with the following parameters: model of amino acids substitution = WAG, discrete gamma model, number of categories =4, tree topology search = NNIs, and optimizing over all other parameters (‘tlr’, proportion of invariant sites estimated from the data).

**HIV-1 dataset.** Sequences of the HIV-1 dataset were sampled from the dataset used by Penn and Stern et al. (2008). Specifically, for this study the amino acids sequences of the genes: *env, gag, nef, pol, rev, tat, vif, vpr* and *vpu* from seven subtypes (A, B, C, D, F, G, J) of HIV-1 group M were used. For each of these genes, a dataset of 50 sequences was composed by collecting all the sequences of the J, G, F and D substrains (32 sequences in total) and randomly sampling six sequences from each of the A, B and C substrains. All datasets were aligned with PRANK using the +F argument. Phylogenetic trees were inferred using PhyML with the following parameters: model of amino acids substitution = WAG, discrete gamma model, number of categories
= 4, tree topology search = best of NNI and SPR, and optimizing over all other parameters.

**Inference of indel parameters using SPARTA**

The SPARTA algorithm was run on all datasets using the following INDELible (Fletcher, Yang 2009) configuration: ‘AMINOACID2’ model, ‘WAG’ substitution model and ‘POW’ indel size model, maximum indel length: 50.

**Algorithm implementation**

The algorithm procedure is described in detail in the results section. The algorithm was implemented in C++ and is freely available at

http://www.tau.ac.il/~talp/supplementary/sparta/sparta.html

For the simulations that are part of the SPARTA algorithm, we have integrated parts of the INDELible source code. Parameters were optimized using an iterative golden search procedure (Press et al. 2002), starting from the root length. To avoid local maxima, three different root length starting points were used: the length of the shortest sequence, the length of the longest sequence, or the number of columns in the MSA. The two other parameters were always searched starting from $IR = 0.075$ and $a$
parameter = 1.55. For the data analyzed in this work, searches were conducted in the following intervals: $IR$ from 0 to 0.16, ‘$a$’ from 1 to 2 and $RL$ from 50 to 1,800.

Acknowledgments

This study was supported by an ISF grant 1092/13 to TP. ELK, HA, and OA are fellows of the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University. Akash Khare and Jessica Albanese explored ideas similar to SPARTA for their undergraduate theses as students of Barrett, the Honors College at Arizona State University.
Literature Cited


Figure Legends

**Figure 1** SPARTA methodology uses Mahalanobis distance to measure the fit of proposed parameters to input data. Presented is a single search step, in which the distance between a proposed set of parameters, $\theta^{(i)}$, and the true unknown parameters $\theta$ is computed. Standard hill-climbing heuristics are used to search for a set of parameters that minimizes the distance between simulated data and input data. The set of parameters that provides the lowest distance is searched using standard hill-climbing heuristics.

**Figure 2** Inference accuracy is positively correlated with the number of simulated MSAs ($N$) used in each search step. Fifty real MSAs were obtained using the “basic parameter configuration” (see Materials and Methods). The parameters of each of these MSAs were then searched for, with different values of $N$. Panel 2A depicts the Mahalanobis distance and the computation time as a function of $N$ and panel 2B shows how each of the inferred parameters depends on $N$. The real parameter values are marked as bold points.

**Figure 3** SPARTA’s inference is better than lambda.pl’s. Fifty real MSAs obtained using the “basic parameter configuration” were given as input to SPARTA as well as to Dawg’s lambda.pl script. The real parameter values are marked as bold points.

**Figure 4** SPARTA’s inference is robust to biases introduced by MSA programs. Fifty sequence datasets obtained using the “basic parameter configuration” were aligned by
either ClustalW, MAFFT or PRANK. The MSAs computed by each alignment program were given as input to SPARTA. The real parameter values are marked as bold points. As reference, we also present the inferred values using the “true” MSAs generated by INDELible.

**Figure 5** SPARTA can be used to simulate MSAs similar to a target MSA. The plot depicts three MSAs. The real Azurin MSA (panel A), a simulated MSA using the parameters the algorithm inferred for the Azurin MSA (IR = 0.0135, \( a = 1.325 \), RL = 119; panel B) and a simulated MSA using INDELible’s default parameters (as described in the Materials and Methods section) (panel C). As the MSA simulated based on the default parameters is 4,242 amino acids long, only the first 200 columns are presented in the plot.

**Figure 6** Distribution of parameter values in real datasets. The algorithm was run on 498 mammalian MSAs obtained from the OrthoMam database as well as 100 COG MSAs. The panels depict the distribution of the inferred parameter values in cases where the p-value was not significant (p > 0.05; 104 OrthoMam genes and 28 COG genes).

**Figure 7** PRANK MSA of the *vif* protein across fifty HIV-1 samples.
Tables

Table 1 Method accuracy for different parameter combinations. Datasets were simulated according to seven alternative parameter configurations. Fifty MSAs were simulated by each configuration and were given as input to SPARTA in order to evaluate its ability to accurately infer the parameter values. Each value is the average inferred parameter and one standard error.

<table>
<thead>
<tr>
<th>Various indel rates</th>
<th>IR = 0.01</th>
<th>IR = 0.02</th>
<th>IR = 0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>RL = 350</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a = 1.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inferred IR</td>
<td>0.016±0.017</td>
<td>0.023±0.006</td>
<td>0.102±0.013</td>
</tr>
<tr>
<td>Inferred a</td>
<td>1.379±0.157</td>
<td>1.324±0.116</td>
<td>1.297±0.061</td>
</tr>
<tr>
<td>Inferred RL</td>
<td>337.1±37.69</td>
<td>341.94±24.76</td>
<td>344.74±23.18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Various slope parameter values</th>
<th>IR = 0.02</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RL = 350</td>
<td>a = 1.1</td>
<td>a = 1.3</td>
<td>a = 1.7</td>
</tr>
<tr>
<td>Inferred IR</td>
<td>0.029±0.018</td>
<td>0.023±0.006</td>
<td>0.021±0.003</td>
</tr>
<tr>
<td>Inferred a</td>
<td>1.261±0.151</td>
<td>1.324±0.116</td>
<td>1.686±0.16</td>
</tr>
<tr>
<td>Inferred RL</td>
<td>319.58±40.19</td>
<td>341.94±24.76</td>
<td>350.36±5.34</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Various root lengths</th>
<th>IR = 0.02</th>
<th>RL = 100</th>
<th>RL = 350</th>
<th>RL = 500</th>
</tr>
</thead>
<tbody>
<tr>
<td>a = 1.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inferred IR</td>
<td>0.021±0.006</td>
<td>0.023±0.006</td>
<td>0.023±0.011</td>
<td></td>
</tr>
<tr>
<td>Inferred a</td>
<td>1.273±0.157</td>
<td>1.324±0.116</td>
<td>1.379±0.093</td>
<td></td>
</tr>
<tr>
<td>Inferred RL</td>
<td>99.14±7.84</td>
<td>341.94±24.76</td>
<td>489.72±25.42</td>
<td></td>
</tr>
</tbody>
</table>

Table 2 Indel dynamics parameters inferred using SPARTA from nine HIV-1 protein MSAs. Each MSA was composed of fifty orthologs.

<table>
<thead>
<tr>
<th>Gene</th>
<th>IR</th>
<th>a</th>
<th>RL</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>env</td>
<td>0.1322</td>
<td>1.55</td>
<td>728</td>
<td>1.00E-14</td>
</tr>
<tr>
<td>gag</td>
<td>0.0195</td>
<td>1.8441</td>
<td>494</td>
<td>0.648083</td>
</tr>
<tr>
<td>nef</td>
<td>0.0195</td>
<td>1.1375</td>
<td>181</td>
<td>0.164597</td>
</tr>
<tr>
<td>pol</td>
<td>0.0872</td>
<td>1.6744</td>
<td>881</td>
<td>1.00E-14</td>
</tr>
<tr>
<td>rev</td>
<td>0.009</td>
<td>1.6586</td>
<td>118</td>
<td>0.997467</td>
</tr>
<tr>
<td>tat</td>
<td>0.0095</td>
<td>1.5756</td>
<td>97</td>
<td>0.997941</td>
</tr>
<tr>
<td>vif</td>
<td>0.0035</td>
<td>1.9014</td>
<td>192</td>
<td>0.989768</td>
</tr>
<tr>
<td>vpr</td>
<td>0.0004</td>
<td>1.55</td>
<td>96</td>
<td>0.999988</td>
</tr>
<tr>
<td>vpu</td>
<td>0.0039</td>
<td>1.2163</td>
<td>451</td>
<td>1.00E-14</td>
</tr>
<tr>
<td>------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>----------</td>
</tr>
</tbody>
</table>

Figures

Figure 1

\[ D(\theta^{(i)}, \theta) \approx D(\mathbf{x}) = \sqrt{(\mathbf{x} - \mu)^T S^{-1} (\mathbf{x} - \mu)} \]
Figure 3
Figure 6

Indel Rate (IR)  
Indel Length Parameter (a)  
Root Length (RL)

Density