SpartaABC: a web server to simulate sequences with indel parameters inferred using an approximate Bayesian computation algorithm

Haim Ashkenazy¹,†, Eli Levy Karin¹,2,†, Zach Mertens³, Reed A Cartwright³,4,* and Tal Pupko¹,†

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

ABSTRACT

Many analyses for the detection of biological phenomena rely on a multiple sequence alignment as input. The results of such analyses are often further studied through parametric bootstrap procedures, using sequence simulators. One of the problems with conducting such simulation studies is that users currently have no means to decide which insertion and deletion (indel) parameters to choose, so that the resulting sequences mimic biological data. Here, we present SpartaABC, a web server that aims to solve this issue. SpartaABC implements an approximate-Bayesian-computation rejection algorithm to infer indel parameters from sequence data. It does so by extracting summary statistics from the input. It then performs numerous sequence simulations under randomly sampled indel parameters. By computing a distance between the summary statistics extracted from the input and each simulation, SpartaABC retains only parameters behind simulations close to the real data. As output, SpartaABC provides point estimates and approximate posterior distributions of the indel parameters. In addition, SpartaABC allows simulating sequences with the inferred indel parameters. To this end, the sequence simulators, Dawg 2.0 and INDELible were integrated. Using SpartaABC we demonstrate the differences in indel dynamics among three protein-coding genes across mammalian orthologs. SpartaABC is freely available for use at http://spartaabc.tau.ac.il/webserver.

INTRODUCTION

Sequence simulation is an extremely important component of phylogenetic studies and many sequence simulators have been previously developed (1–12). The tasks for which sequence simulations are used vary greatly and span a wide range of scientific questions. For example, Worobey et al. used sequence simulations to investigate the origins of influenza A virus within and between hosts (13). Shapiro et al. utilized sequence simulations in their study of early events of ecological differentiation of bacterial genomes (14). Gossmann and Schmid included sequence simulations as part of their analysis of post-duplication selective forces on genes in Arabidopsis thaliana (15). Sequence simulators are also often used in studies that aim to evaluate the performance of alignment and tree reconstruction algorithms (16–23). Finally, sequence simulations are an integral part of parametric bootstrap test procedures, which are used, for example, to test for the constancy of evolutionary rates (24), to study the fit of various evolutionary models to real sequences (25,26), to detect traits that impact the rate of evolution (27,28) and to compare competing tree topologies (29–31).

Sequence simulators provide in-silico generated datasets under different evolutionary scenarios. The complete evolutionary process relies on a substitution model (e.g. 32–38) as well as a model of insertion and deletion. The occurrence of indel events is defined relative to the substitution process and is controlled by the IR parameter—the indel-to-substitution rate ratio. The length of the indel is often modeled using a power-law distribution, controlled by its shape parameter ‘A’. This distribution is characterized by a reverse relationship between an indel size and its probability. Finally, the root length parameter RL controls the length of the sequence at the root of the tree (the start of evolution).
the simulation). Although the root length is not a pure indel parameter, it strongly affects the resulting multiple sequence alignment (MSA). Until recently, no methodology was available for users to determine the values of these parameters in a way that best reflects the indel dynamics in their datasets of interest.

We recently developed the SpartaABC algorithm (Levy Karin, Shkedy et al. submitted for publication), an approximate Bayesian computation rejection algorithm to infer indel parameters from sequence data. SpartaABC focuses on the inference of the above-mentioned three indel parameters. To this end, SpartaABC extracts a vector of summary statistics from its input; it then performs repeated simulations using an integrated sequence simulator (8,12) under various indel parameters. From each such simulated dataset it extracts a vector of summary statistics and computes its distance from the vector extracted for the input using a weighted Euclidean distance. SpartaABC retains a subset of the simulations for which the distance from the input was small enough. The parameter sets from simulations with a small distance are used to estimate the indel parameters behind the input. Using a simulation study, the SpartaABC algorithm was shown to accurately infer indel parameter values under various conditions (Levy Karin, Shkedy et al. submitted for publication). Thus, sequences simulated using the SpartaABC inferred indel parameters resemble the input data in terms of their indel properties, much more so, than when sequence simulators are run with default parameters.

Here we use the SpartaABC algorithm as part of a broader web service, which provides the following: (i) MSA reconstruction (optional), (ii) tree reconstruction (optional), (iii) inference of indel dynamics and (iv) sequence simulation based on the inferred indel parameters (optional). Visual and textual outputs of these services are offered as downloadable files.

MATERIALS AND METHODS

Input

The SpartaABC web server requires sequence data (either nucleotide or protein) as input. The user can provide either an MSA or a set of unaligned sequences. If unaligned data are provided, the user will be asked to choose between the programs MAFFT (39,40) and PRANK (41) to align them. An optional input to the SpartaABC web server is a phylogenetic tree. If the user does not provide a phylogenetic tree, the maximum likelihood tree will be computed based on the MSA of the sequences, using RAxML (42). SpartaABC integrates two sequence simulators: Dawg 2.0 (12), which is the default, and INDELible (8). Finally, the user can indicate the number of simulated datasets to produce based on the indel parameters inferred from the input. An illustration of the computational stages performed by the SpartaABC web server is presented in Figure 1.

Summary statistics

The summary statistics computed by SpartaABC are detailed in the OVERVIEW section of the web server. Among them are the average gap length, the total number of gaps and the MSA length. Based on the summary statistics extracted from the input MSA and each simulation, SpartaABC computes a weighted Euclidean distance. The weights used by SpartaABC are also available for download from the SOURCE & USAGE section of the web server. In addition, the extracted summary statistics values from the input MSA and each simulation in the SpartaABC run are available for download.

Indel parameters search space

Throughout its computation, SpartaABC proposes 100,000 indel parameter combinations by sampling values of each of the parameters from a prior uniform distribution. Specifically, the ‘A’ parameter value is sampled from a wide range: (1, 2]; the IR parameter value is sampled by default from the range: [0, 0.05], but this range can be extended by the user up to [0, 0.1]. Finally, the RL parameter range is determined empirically according to the input provided by the user. Let L denote the longest sequence in the user-provided input, then the search range of the RL parameter is [50, 1.1 × L].

Output

SpartaABC provides a step-by-step progress report and estimation of the expected run time. Upon completion of the SpartaABC computation, all examined indel parameter combinations and their distance from the input dataset are available to the user as a downloadable file. Out of these, the 50 parameter combinations with the smallest distance are used to approximate the posterior distributions of the indel parameters. These distributions are presented to the user in three plots, where the x-axis is the entire search range of each indel parameter and the y-axis is the density. An example for such plots is given in Figure 2. In addition, SpartaABC computes the posterior expectations based on the inferred posterior distributions to yield point estimates of the indel parameters. As its final step, the web server simulates datasets using the indel parameters point estimates, according to the number of replicates determined by the user. The substitution model and parameters used in the sequence simulation step are estimated and selected according to the AICc (43) criterion by jModelTest (44) or protTest (45), for DNA or protein input, respectively. The user can download a zipped file of these simulated datasets as well as the sequence simulator control file. Finally, the MSA and phylogenetic tree from which SpartaABC inferred the indel parameters are presented visually using Wasabi (46).

Implementation

The SpartaABC web server runs on a Linux cluster of 2.6 GHz AMD Opteron processors, equipped with 4 GB RAM per quad-core node. The server runs up-to-date versions of the supported multiple alignment and tree reconstruction programs. The SpartaABC algorithm was implemented in C++. We provide its source code, a precompiled version for UNIX systems, a short manual and a run example in the SOURCE & USAGE section of the web server. In addition, the web server contains a frequently asked questions page to provide additional information concerning the algorithm and methodology.
CASE STUDY

SERPINA7, PTH1R and CFTR are genes known to play a role in the human diseases: thyroxine-binding globulin deficiency, chondrodysplasia and cystic fibrosis, respectively (47–49). In order to examine their indel dynamics, we obtained their sets of unaligned coding sequences across >30 mammalian orthologous species from the OrthoMam database (50). These datasets are available for download at the GALLERY section of the web server. We then analyzed each of these sets using the SpartaABC web server. First, an MSA was computed for each unaligned set of sequences using the server’s default MSA program, MAFFT (40). Second, a phylogenetic tree was reconstructed using RAxML (42). Third, the MSA and the tree were used to infer indel parameters. We found, that in spite of the fact that all three analyzed coding sequences are involved in human diseases and have roughly the same number of mammalian orthologs, they display substantially different indel dynamics (Figure 2). Specifically, the IR parameter inferred for SERPINA7 is 5-fold smaller than that inferred for CFTR, with PTH1R taking an IR value in between the other two. In addition, the inferred RL parameters corresponded to the different lengths of the genes. Finally, all genes displayed a tendency for longer indels as evident by their inferred ‘A’ parameter. All three inferred ‘A’ values were close to 1.0, yielding power low distributions where longer indels are more probable compared to power low distributions with a high ‘A’ value. From these results we conclude that even when examining orthologous genes within the same taxonomic class and similar biological contexts, it is important to characterize the indel dynamics of each gene individually in order to best mimic biological data.

In the following example, we demonstrate the utility of the SpartaABC web server to test specific evolutionary hypotheses using a parametric bootstrap procedure, in which the sequences are generated based on the indel parameters inferred from the data. Here, we focused on the comparison of the indel dynamics between the coding region of SERPINA7 (as analyzed above) to the entire SERPINA7 gene (exons and introns included). To this end, we obtained the full SERPINA7 gene sequences across 35 mammalian orthologs from the ENSEMBL database (51). Using SpartaABC web server, we found that over the whole gene, the IR parameter is much higher (0.0166) compared to that inferred in the coding sequence only (0.004), suggesting that...
indels are much more frequent when intronic regions are included in the analysis compared to examining only exonic regions. A much smaller difference was found in the inferred ‘A’ parameter (1.036 and 1.2 for the full and coding-only SERPINA7, respectively), suggesting that the main difference between the full and coding-only SERPINA7 is the frequency of indels, rather than their size. We hypothesized that such a difference may stem from selection against the introduction of indels in a specific region that resides within the exons of the analyzed gene. To statistically test this hypothesis, we first measured the longest stretch of consecutive columns without gap characters in the MSA of the full SERPINA7 gene. We found that this stretch was 174 columns in length, which reside within the second human exon of this gene (starting at position 3726 of the MSA). Using simulations which do not prefer one sequence position over the other for indel events, we could test how likely it is to observe a gap-free stretch of consecutive columns of such length. We thus compared the length of the SERPINA7 stretch to those computed from 100 simulated MSAs produced by the SpartaABC web server according to the SERPINA7 inferred indel parameters using Dawg 2.0 (12). In all 100 simulated datasets we found that the longest stretch without any gap characters did not exceed 85 columns in length, suggesting the SERPINA7 stretch is significantly longer than one could expect (empirical P-value < 0.01). The data (e.g., MSAs and trees) and the analyses associated with this example are provided in the GALLERY section of the web server. In conclusion, indel dynamics can vary along a specific gene and using sequence simulations it is possible to detect gene regions that deviate from the average indel dynamics inferred for the entire sequence.

FUNDING
BSF [2015247 to R.A.C., T.P.]; ISF [1092/13, 802/16 to T.P.]; Edmond J. Safra Center for Bioinformatics at Tel Aviv University Fellowship (to E.L.K.). Funding for open access charge: BSF [2015247 to R.A.C., T.P.]; ISF [1092/13, 802/16 to T.P.].

Conflict of interest statement. None declared.

REFERENCES