README

empiricIST: An Integrated Software and analysis Tool for analyzing time-sampled sequence data such as EMPIRIC

Inês Fragata, Sebastian Matuszewski, Jeffrey D. Jensen and Claudia Bank

September 4, 2017

Contents

1 Introduction

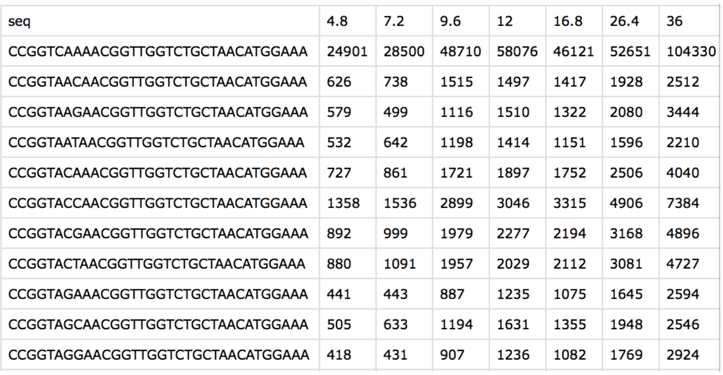
This program serves as an extension of the Bayesian Monte Carlo Markov chain (MCMC) method described in Bank et al. (2014) for estimating selection coefficients (growth rates) from engineered-mutation-driven experimental evolution data. These data are based on methods – such as EMPIRIC – in which specific mutations are engineered, introduced and compared against each other and a reference (e.g., wild-type) sequence. All mutants (and the wild-type) are assumed to have evolved together but independently in bulk competition over a number of generations with samples taken throughout the course of the experiment. Growth rate estimates are obtained from the number of reads obtained from deep sequencing. The motivation for the *empiricIST* software package is to provide an integrative software tool for the analysis of deep-mutational scanning data, and includes separate programs for processing raw sequence data, growth rate estimation, post-processing and analyses of growth rate estimates. Before using the software, please read the two accompanying papers by Bank et al. (2014) and **?** that describe the methods and their underlying assumptions in more detail.

2 Input data

`empiricIST_MCMC` input data is expected to be in csv-format (i.e., comma-separated values) following a specific ordering and UNIX line endings. As part of the *empiricIST* software package, however, we provide a python script – `empiricIST_MCMC_Input.py` – that adjusts the raw data to match the specific input format needed. A minimal example of the raw data that is required to generate the `empiricIST_MCMC` input file is shown in Figure 1. Note that the raw data itself also needs to be csv-formatted with column entries separated by a comma (','), a semi-colon (';') or a tab ('\t'). Furthermore, the raw data file needs to have (exactly) one column either called 'sequence', 'Sequence', 'seq' or 'Seq' with at least two rows (the wild-type reference in the first row and a mutant) and at least additional 3 columns corresponding to the number of sequencing reads per sampled time point (and header cells giving the time of sampling).

2.1 `empiricIST_MCMC_Input`.py

The program `empiricIST_MCMC_Input.py` is written in Python (2.7) and serves as a link between the (raw) time-sampled sequence data (e.g., obtained from a deep-mutational scanning experiment) and the `empiricIST_MCMC` simulation program



seq	4.8	7.2	9.6	12	16.8	26.4	36
CCGGTCAAAACGGTTGGTCTGCTAACATGGAAA	24901	28500	48710	58076	46121	52651	104330
CCGGTAACAACGGTTGGTCTGCTAACATGGAAA	626	738	1515	1497	1417	1928	2512
CCGGTAAGAACGGTTGGTCTGCTAACATGGAAA	579	499	1116	1510	1322	2080	3444
CCGGTAATAACGGTTGGTCTGCTAACATGGAAA	532	642	1198	1414	1151	1596	2210
CCGGTACAAACGGTTGGTCTGCTAACATGGAAA	727	861	1721	1897	1752	2506	4040
CCGGTACCAACGGTTGGTCTGCTAACATGGAAA	1358	1536	2899	3046	3315	4906	7384
CCGGTACGAACGGTTGGTCTGCTAACATGGAAA	892	999	1979	2277	2194	3168	4896
CCGGTACTAACGGTTGGTCTGCTAACATGGAAA	880	1091	1957	2029	2112	3081	4727
CCGGTAGAAACGGTTGGTCTGCTAACATGGAAA	441	443	887	1235	1075	1645	2594
CCGGTAGCAACGGTTGGTCTGCTAACATGGAAA	505	633	1194	1631	1355	1948	2546
CCGGTAGGAACGGTTGGTCTGCTAACATGGAAA	418	431	907	1236	1082	1769	2924

Figure 1 – Schematic illustration of the minimal data needed to run the `empiricIST_MCMC` program.

for the estimation of mutant growth rates 'r'. While it primarily ensures that the input data matches the input format required by the MCMC simulation program, it comes with additional options that will be detailed here.

The general usage is as follows: After opening a command-line interface (e.g., Shell, Terminal) and navigating to the location of the `empiricIST_MCMC_Input.py` file, the program can be executed by typing

```
python empiricIST_MCMC_Input.py [options] .
```

Note that this requires that the 'PATHVARIABLE' for Python has been set correctly on your system. Please consult the online Python documentation for further details (https://docs.python.org/2/). Without specifying any options the program will exit with an error and provide a short documentation on its usage. To execute the program, the name of the (raw) data input file and the start of the sequencing read data needs to be specified (by invoking the '-f' and '-s' option, respectively). All options and their usage are given in Table 1.

Table 1 – A summary of the options of the `empiricIST_MCMC_Input.py` program.

Short/Long option	**Accepted values**	**Description**
-h, –help	none	When the '-h'-option is invoked, a short documentation on the usage of the program is shown. Note that, if this option is invoked, the python program is not executed.
-f, –file=	string	The '-f'-option is a mandatory option, which passes the name of the (raw) data input file (csv formatted) to the python program. Files created by the python program will take the name of the input file and add option-dependent specific file identifiers.
-s, –skipcol=	integer	The '-s'-option is a mandatory option, which takes an integer value corresponding to the number of descriptive columns that precede the actual 'data matrix' of sequencing read counts. For example, for the raw data depicted in Figure 1, the user would have to pass '-s 1' (or equivalently '–skipcol=1'). Please note that the data matrix must always span all remaining columns.

-o, –outlier=	'detect' or 'impute'	When the '-o'-option is invoked, the python program will perform an outlier analysis. If '-o detect' (or equivalently '–outlier=detect'), the python program performs a log-linear regression analysis for all mutants. Data points are then classified as outliers on the basis of the DFBeta statistic with a cut-off value of 2 (with data points surpassing this cut-off regarded as outliers). For more details please consult Bank et al. (2014). If '-o impute' (or equivalently '–outlier=impute'), the python program performs a log-linear regression analysis for all mutants. Data points are then classified as outliers on the basis of the DFBeta statistic with a cut-off value of 2 and their studentized residuals with a cut-off value of 3 (with data points surpassing *both* cut-offs are regarded as outliers) and imputed as described in **?**. Furthermore, an additional output file – whose name consists of 'ImputedData' along with the input file name – is produced that lists all imputed data points. In particular, the output is a simple matrix where rows denote different mutants and columns correspond to the different time points. An entry of '1' indicates that the data has *not* been imputed; an entry of '0' indicates that the data point has been imputed.
-l, –leadseq=	integer	When the '-l'-option is invoked, the first 'integer' characters, that precede the original mutant sequence (e.g., sites that function as DNA barcode or sequence tag), are removed.
-t, –trailseq=	integer	When the '-l'-option is invoked, the first 'integer' characters, that trail the original mutant sequence (e.g., sites that function as DNA barcode or sequence tag), are removed.
-p, –pool	none	When the '-p'-option is invoked, the DNA-sequences (characterizing the different mutants) are translated into amino acids. The data is then pooled based on their amino acid sequence, assuming that identical amino acid sequences, though differing in their DNA sequence (synonymous mutants), have the same growth rate (but different initial population sizes). Note that even if the '-p'-option is not invoked, data is pooled based on the sequence name (which can be any string and not only letters from the DNA alphabet).
-e, –exp	none	When the '-e'-option is invoked, time points are taken to be in hours instead of generations (i.e., the default).
-g, –group=	integer	When the '-g'-option is invoked, the data is grouped into subsets of mutants each of minimal size 'integer'. This results in more data sets with less mutants, such that the per-data set computation time is reduced, without affecting parameter estimates or the shape of the log-likelihood surface (compared to analysis of the full data set). The program also ensures, that mutants with identical mutant or protein ID (i.e., mutants that have an identical DNA- or amino acid sequence) remain in the same data sub-set as they are assumed to evolve under an identical growth rate (r). The name of the output file (i.e., the MCMC input file) is composed of the standard output file identifier 'MCMCInput', the grouping identifier (a consecutive number of the sub data sets created) and the name of the input file.

-i, –initialize	none	When the '-i'-option is invoked, and additional input file is created that specifies the initial growth rates (r) and initial population sizes (c) for all mutants based on the log-linear regression. Note that this could potentially bias the MCMC algorithm, since the initial starting point of the Markov chain could be trapped in a local log-likelihood optimum. Often, however, the median of the growth rates and the initial population sizes from the MCMC-DFE simulations are close enough to the corresponding estimates from the log-linear regression such that starting at these values could shorten the burn-in period and, thus, reduce the run time. For mutants with identical DNA or amino acid sequence, the mean initial population size is calculated (from these mutants) and taken as the starting value for the MCMC simulation program. Given the estimated mean initial population size, the growth rate is estimated from the log-linear regression. The name of the output file is composed of the standard output file identifier 'MCMCInput', an optional grouping identifier (see '-g'-option), the name of the input file, and an initialization file identifier '`_inputRC`'.

An illustration of the output file produced by the python program (i.e., the input file for the MCMC simulation program) is depicted in Figure 2.

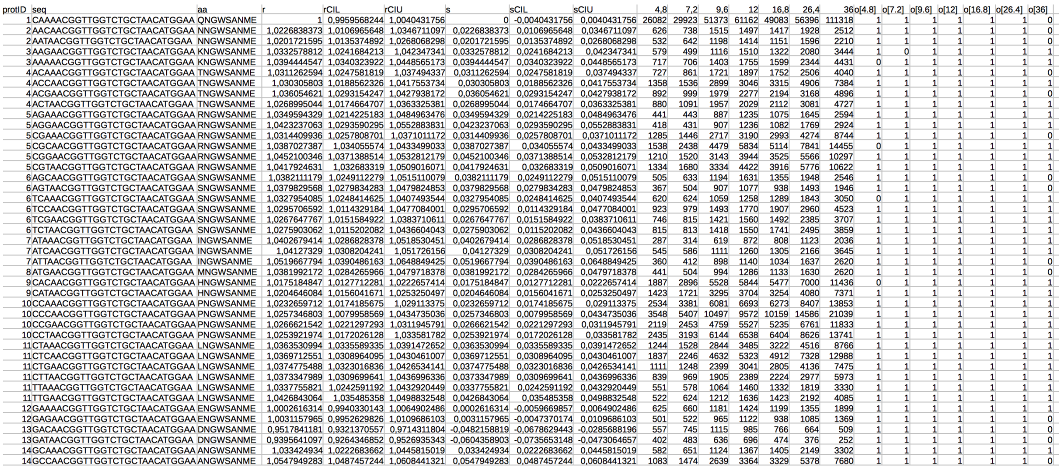
Depending on the invoked options, the name of the MCMC simulation input file that is produced by the python program is given by the standard output file identifier 'MCMCInput', an optional grouping identifier (see '-g'-option) and the name of the raw data file.

3 Usage

We provide executables for Mac OS X, Windows and Linux. The C++ source code along with a system specific makefile are provided under a GNU General Public License as published by the Free Software Foundation. If you do not need to compile the program yourself you can skip the next subsection.

3.1 Compilation

Note that compilation requires that the Gnu Scientific Library (gsl-library) is installed on your system. Information on how to install the gsl-library can be found under `http://www.gnu.org/software/gsl/`. On Windows the easiest way to obtain the gsl-library is to install Cygwin (`http://www.cygwin.com/`) including the developers (all) packages. Alternatively, MinGW (`http://www.mingw.org/`) provides a ”minimalist GNU for Windows” development environment. Under MinGW



protID	seq	aa	r	rCIL	rCIU	s	sCIL	sCIU	4,8	7,2	9,6	12	16,8	26,4	36	o[4.8]	o[7.2]	o[9.6]	o[12]	o[16.8]	o[26.4]	o[36]
1	CAAAACGGTTGGTCTGCTAACATGGAA	QNGWSANME	1	0,9959568244	1,0040431756	0	-0,0040431756	0,0040431756	26082	29923	51373	61162	49083	56396	111318	1	1	1	1	1	1	0
2	AACAACGGTTGGTCTGCTAACATGGAA	NNGWSANME	1,0226838373	1,0106965648	1,0346711097	0,0226838373	0,0106965648	0,0346711097	626	738	1515	1497	1417	1928	2512	1	1	1	1	1	1	0
2	AATAACGGTTGGTCTGCTAACATGGAA	NNGWSANME	1,0201721595	1,0135374892	1,0268068298	0,0201721595	0,0135374892	0,0268068298	532	642	1198	1414	1151	1596	2210	1	1	1	1	1	1	0
3	AAGAACGGTTGGTCTGCTAACATGGAA	KNGWSANME	1,0332578812	1,0241684213	1,042347341	0,0332578812	0,0241684213	0,042347341	579	499	1116	1510	1322	2080	3444	1	0	1	1	1	1	1
3	AAAAACGGTTGGTCTGCTAACATGGAA	KNGWSANME	1,0394444547	1,0340323922	1,0448565173	0,0394444547	0,0340323922	0,0448565173	717	706	1403	1755	1599	2344	4431	0	1	1	1	1	1	1
4	ACAAACGGTTGGTCTGCTAACATGGAA	TNGWSANME	1,0311262594	1,0247581819	1,037494337	0,0311262594	0,0247581819	0,037494337	727	861	1721	1897	1752	2506	4040	1	1	1	1	1	1	0
4	ACCAACGGTTGGTCTGCTAACATGGAA	TNGWSANME	1,030305803	1,0188562326	1,0417553734	0,030305803	0,0188562326	0,0417553734	1358	1536	2899	3046	3315	4906	7384	1	1	1	1	1	1	1
4	ACGAACGGTTGGTCTGCTAACATGGAA	TNGWSANME	1,036054621	1,0293154247	1,0427938172	0,036054621	0,0293154247	0,0427938172	892	999	1979	2277	2194	3168	4896	1	1	1	1	1	1	0
4	ACTAACGGTTGGTCTGCTAACATGGAA	TNGWSANME	1,0268995044	1,0174664707	1,0363325381	0,0268995044	0,0174664707	0,0363325381	880	1091	1957	2029	2112	3081	4727	1	1	1	1	1	1	1
5	AGAAACGGTTGGTCTGCTAACATGGAA	RNGWSANME	1,0349594329	1,0214225183	1,0484963476	0,0349594329	0,0214225183	0,0484963476	441	443	887	1235	1075	1645	2594	1	1	1	1	1	1	1
5	AGGAACGGTTGGTCTGCTAACATGGAA	RNGWSANME	1,0423237063	1,0293590295	1,0552883831	0,0423237063	0,0293590295	0,0552883831	418	431	907	1236	1082	1769	2924	1	1	1	1	1	1	1
5	CGAAACGGTTGGTCTGCTAACATGGAA	RNGWSANME	1,0314409936	1,0257808701	1,0371011172	0,0314409936	0,0257808701	0,0371011172	1285	1446	2717	3190	2993	4274	8744	1	1	1	1	1	1	0
5	CGCAACGGTTGGTCTGCTAACATGGAA	RNGWSANME	1,0387027387	1,034055574	1,0433499033	0,0387027387	0,034055574	0,0433499033	1538	2438	4479	5834	5114	7841	14455	0	1	1	1	1	1	1
5	CGGAACGGTTGGTCTGCTAACATGGAA	RNGWSANME	1,0452100346	1,0371388514	1,0532812179	0,0452100346	0,0371388514	0,0532812179	1210	1520	3143	3944	3525	5566	10297	1	1	1	1	1	1	1
5	CGTAACGGTTGGTCTGCTAACATGGAA	RNGWSANME	1,0417924631	1,032683319	1,0509016071	0,0417924631	0,032683319	0,0509016071	1334	1680	3334	4422	3916	5776	10622	1	1	1	1	1	1	1
6	AGCAACGGTTGGTCTGCTAACATGGAA	SNGWSANME	1,0382111179	1,0249112279	1,0515110079	0,0382111179	0,0249112279	0,0515110079	505	633	1194	1631	1355	1948	2546	1	1	1	1	1	1	0
6	AGTAACGGTTGGTCTGCTAACATGGAA	SNGWSANME	1,0379829568	1,0279834283	1,0479824853	0,0379829568	0,0279834283	0,0479824853	367	504	907	1077	938	1493	1946	1	1	1	1	1	1	0
6	TCAAACGGTTGGTCTGCTAACATGGAA	SNGWSANME	1,0327954085	1,0248414625	1,0407493544	0,0327954085	0,0248414625	0,0407493544	620	624	1059	1258	1289	1843	3050	0	1	1	1	1	1	1
6	TCCAACGGTTGGTCTGCTAACATGGAA	SNGWSANME	1,0295706592	1,0114329184	1,0477084001	0,0295706592	0,0114329184	0,0477084001	923	979	1493	1770	1907	2960	4523	1	1	1	1	1	1	1
6	TCGAACGGTTGGTCTGCTAACATGGAA	SNGWSANME	1,0267647767	1,0151584922	1,0383710611	0,0267647767	0,0151584922	0,0383710611	746	815	1421	1560	1492	2385	3707	1	1	1	1	1	1	1
6	TCTAACGGTTGGTCTGCTAACATGGAA	SNGWSANME	1,0275903062	1,0115202082	1,0436604043	0,0275903062	0,0115202082	0,0436604043	815	813	1418	1550	1741	2495	3859	1	1	1	1	1	1	1
7	ATAAACGGTTGGTCTGCTAACATGGAA	INGWSANME	1,0402679414	1,0286828378	1,0518530451	0,0402679414	0,0286828378	0,0518530451	287	314	619	872	808	1123	2036	1	1	1	1	1	1	1
7	ATCAACGGTTGGTCTGCTAACATGGAA	INGWSANME	1,04127329	1,0308204241	1,051726156	0,04127329	0,0308204241	0,051726156	545	586	1111	1260	1305	2166	3645	1	1	1	1	1	1	1
7	ATTAACGGTTGGTCTGCTAACATGGAA	INGWSANME	1,0519667794	1,0390486163	1,0648849425	0,0519667794	0,0390486163	0,0648849425	360	412	898	1140	1034	1637	2620	1	1	1	1	1	1	0
8	ATGAACGGTTGGTCTGCTAACATGGAA	MNGWSANME	1,0381992172	1,0284265966	1,0479718378	0,0381992172	0,0284265966	0,0479718378	441	504	994	1286	1133	1630	2620	1	1	1	1	1	1	0
9	CACAACGGTTGGTCTGCTAACATGGAA	HNGWSANME	1,0175184847	1,0127712281	1,0222657414	0,0175184847	0,0127712281	0,0222657414	1887	2896	5528	5844	5477	7000	11436	0	1	1	1	1	1	1
9	CATAACGGTTGGTCTGCTAACATGGAA	HNGWSANME	1,0204646084	1,0156041671	1,0253250497	0,0204646084	0,0156041671	0,0253250497	1423	1721	3295	3704	3254	4080	7371	1	1	1	1	1	1	1
10	CCAAACGGTTGGTCTGCTAACATGGAA	PNGWSANME	1,0232659712	1,0174185675	1,029113375	0,0232659712	0,0174185675	0,029113375	2534	3381	6081	6693	6273	8407	13853	1	1	1	1	1	1	1
10	CCCAACGGTTGGTCTGCTAACATGGAA	PNGWSANME	1,0257346803	1,0079958569	1,0434735036	0,0257346803	0,0079958569	0,0434735036	3548	5407	10497	9572	10159	14586	21039	1	1	1	1	1	1	1
10	CCGAACGGTTGGTCTGCTAACATGGAA	PNGWSANME	1,0266621542	1,0221297293	1,0311945791	0,0266621542	0,0221297293	0,0311945791	2119	2453	4759	5527	5235	6761	11833	1	1	1	1	1	1	1
10	CCTAACGGTTGGTCTGCTAACATGGAA	PNGWSANME	1,0253921974	1,0172026128	1,033581782	0,0253921974	0,0172026128	0,033581782	2435	3193	6144	6538	6404	8626	13741	1	1	1	1	1	1	1
11	CTAAACGGTTGGTCTGCTAACATGGAA	LNGWSANME	1,0363530994	1,0335589335	1,0391472652	0,0363530994	0,0335589335	0,0391472652	1244	1528	2844	3485	3222	4516	8766	1	1	1	1	1	1	1
11	CTCAACGGTTGGTCTGCTAACATGGAA	LNGWSANME	1,0369712551	1,0308964095	1,0430461007	0,0369712551	0,0308964095	0,0430461007	1837	2246	4632	5323	4912	7328	12988	1	1	1	1	1	1	1
11	CTGAACGGTTGGTCTGCTAACATGGAA	LNGWSANME	1,0374775488	1,0323016836	1,0426534141	0,0374775488	0,0323016836	0,0426534141	1111	1248	2399	3041	2805	4136	7475	1	1	1	1	1	1	1
11	CTTAACGGTTGGTCTGCTAACATGGAA	LNGWSANME	1,0373347989	1,0309699641	1,0436996336	0,0373347989	0,0309699641	0,0436996336	839	969	1905	2389	2224	2977	5973	1	1	1	1	1	1	1
11	TTAAACGGTTGGTCTGCTAACATGGAA	LNGWSANME	1,0337755821	1,0242591192	1,0432920449	0,0337755821	0,0242591192	0,0432920449	551	578	1064	1460	1332	1819	3330	1	1	1	1	1	1	1
11	TTGAACGGTTGGTCTGCTAACATGGAA	LNGWSANME	1,0426843064	1,035485358	1,0498832548	0,0426843064	0,035485358	0,0498832548	522	624	1212	1636	1423	2192	4085	1	1	1	1	1	1	1
12	GAAAACGGTTGGTCTGCTAACATGGAA	ENGWSANME	1,0002616314	0,9940330143	1,0064902486	0,0002616314	-0,0059669857	0,0064902486	625	660	1181	1424	1199	1355	1899	1	1	1	1	1	1	1
12	GAGAACGGTTGGTCTGCTAACATGGAA	ENGWSANME	1,0031157965	0,9952629826	1,0109686103	0,0031157965	-0,0047370174	0,0109686103	501	522	965	1122	938	1085	1369	1	1	1	1	1	1	0
13	GACAACGGTTGGTCTGCTAACATGGAA	DNGWSANME	0,9517841181	0,9321370557	0,9714311804	-0,0482158819	-0,0678629443	-0,0285688196	557	745	1115	985	766	664	509	1	1	1	1	1	1	0
13	GATAACGGTTGGTCTGCTAACATGGAA	DNGWSANME	0,9395641097	0,9264346852	0,9526935343	-0,0604358903	-0,0735653148	-0,0473064657	402	483	636	696	474	376	252	1	1	1	1	1	1	0
14	GCAAACGGTTGGTCTGCTAACATGGAA	ANGWSANME	1,033424934	1,0222683662	1,0445815019	0,033424934	0,0222683662	0,0445815019	582	651	1124	1367	1405	2149	3302	1	1	1	1	1	1	1
14	GCCAACGGTTGGTCTGCTAACATGGAA	ANGWSANME	1,0547949283	1,0487457244	1,0608441321	0,0547949283	0,0487457244	0,0608441321	1083	1474	2639	3364	3329	5378	7680	1	1	1	1	1	1	0

Figure 2 – Schematic illustration of the output produced by the python program (which serves as input for the MCMC simulation program). This input data has been created using the minimal raw data shown in Fig. 1, where the first five bases and the last base have been discarded (barcodes; '-l' and '-t' option). DNA sequences ('seq' column) have been translated to amino acids ('aa' column) and pooled ('-p' option), such that identical amino acid sequences have the same protein ID ('protID' column). Estimates of the growth rates 'r' ('r' column) and the selection coefficient 's' ('s' column) along with the 95%-confidence intervals are based on the log-linear regression (where 'rCIL' and 'rCIU' give the lower and upper boundary of the confidence interval for the growth rate r, respectively. Notation is analogous for the selection coefficient 's'.). Please note, that while the MCMC simulation program assumes that mutants with identical sequence information (i.e., for sequences with identical 'protID') evolve at the same growth rate, log-linear estimates for 'r' and 's' are based on individual mutants. The columns '4.8', '7.2', '9.6', '12', '16.8', and '26.4' give the number of sequencing reads obtained from sampling at these time points for each mutant (row). If the '-o detect'-option is invoked, the matrix of sequencing reads is followed by an outlier matrix for the corresponding time points and mutants, where '0' indicate data points that where classified as outliers.

though, the gsl-library needs to be installed independently. A short instruction is given in the following paragraph.

Compilation of the program has successfully been tested on MacOSX (10.12.6) using 'clang' (version 6.0) and 'gcc' (version 4.9), on Ubuntu (14.04 LTS) using 'g++' (version 4.8.2), and on Windows (8.1) using 'g++' (version 4.8.1 under MinGW; version 4.9.2 under Cygwin).

Windows

Compilation of the program under Windows either requires Cygwin – i.e., a large collection of GNU and open source tools, which provide functionality similar to a Linux distribution on Windows – or MinGW – a complete open source programming tool set for the development of native Windows applications including both different compilers and a "minimal system" bourne shell command line interpreter system MSYS.

The easiest way to compile the program is by using the provided makefile. For that simply type

```
make
```

after having navigated to the folder where the source code is stored.

Please note that when using Cygwin commands and folder navigation are different to those from a native Windows Sehll (e.g., Powershell) For example, under Cygwin you first need to type

```
cd /cygdrive/c
```

to navigate to the drive 'C:\'. Furthermore, note that Cygwin does not allow to link libraries statically.

When using MinGW/MSYS please check before compilation whether the gsl-library is installed. If not, a rough manual on how this can be done is given below. When already installed, the easiest way to compile the program is again to use the provided makefile. For that simply type

```
make
```

after having navigated to the folder where the source code is stored. Note that to navigate to the drive 'C:\' under MSYS one first has to type

```
cd /c .
```

Compilation and installation of the gsl-library from source Before compilation and installation of the gsl-library please read and follow the installation instructions provided with the most recent version of gsl.

1. Download the latest version of gsl from `http://ftpmirror.gnu.org/gsl/`
2. Navigate to the place where the downloaded tar archive is stored and unpack it by typing

   ```
   tar -zxvf gsl-x.xx.tar.gz ,
   ```

   where 'x.xx' should be replaced by the version number.
3. Navigate to 'gsl-x.xx/' and carefully read the 'INSTALL' document and follow the instructions to configure, make and install the gsl-library:

   ```
   ./configure
   make
   make install
   ```

4. GSL binaries, headers and library files are installed automatically in the 'bin/', 'include/gsl/', and 'lib/' subdirectories (if not specified otherwise; in that case you would also need to adjust the linker and compiler flags in the makefile).

Linux and MacOS X

The easiest way to compile the program is by using the provided makefile. For that simply use the shell and type

```
make
```

after having navigated to the folder where the source code is stored. Please note that you might want to adjust the makefile, in particular, to change the name of the executable which is by default set to '*empiricIST*_`MCMC.out`'.

3.2 Execution

Note that the *empiricIST*_`MCMC` program is a command line program which is run from a command-line interface (e.g., Shell, Terminal, Powershell), with arguments and parameters being passed over the command line. To ease execution of the program we provide OS specific scripts, where parameters and options can be specified by the user. An overview and description of the parameters and options can be found in Table 2. Examples on how to manually execute the program are given when running the program without any parameters.

Under Windows there are two different scripts for running the program, depending on whether the program is run under Cygwin or using Powershell. Note that when using Cygwin, the source code needs to be compiled first (since Cygwin does not allow to compile statically) and 'PathToDataFile' needs to be preceded by '/cygdrive/c/' (assuming that the datafile is stored on the 'c' drive).

Table 2 – A summary of the parameters of the MCMC program.

Category	**Accepted values**	**Description**
Parameter		
Data		
-file	string	MANDATORY: Give the full path to the datafile (e.g., /users/me/PathToData/reads.csv).
-prefix	string	The name of the output file prefix (default DFE). The program will take the name of the input file and add the 'prefix', the time of execution and the file identifier (e.g., '`_C`'). This produces for example /PathToData/`_'file'_'prefix'_'timeStamp'_C.txt`.
-skipCol	integer, $ ≥0	MANDATORY: Number of columns to skip in data file before read numbers start.
-outliers		Pass the -outlier option if there is an outlier matrix in the data file.
MCMC		
-burnin	integer, $ ≥ 0	Number of accepted values that are discarded (burn-in period). During the burn-in period the parameters of the proposal distribution are optimized. By default 100,000 is used.
-subsampling	integer, $ ≥ 0	After the burn-in period only every 'subsampling' accepted value is recorded (i.e., written to file). By default 1,000 is used.
-noSets	integer, $ 0	Number of output data sets that are recorded each of size 'set'. By default 10 is used.
-set	integer, $ 0	Number of recorded samples per set. By default 1,000 is used.
-growthRateSD	double, $ > 0	Standard deviation of the proposal distribution of growth rates 'r' drawn from a Gaussian distribution. By default 0.0004 is used.
-popSizeSD	double, $ > 0	Scale parameter of the proposal distribution of initial population sizes 'c' drawn from a Cauchy distribution. By default 0.0002 is used.
-initital	string	Specify the file (including its path) to the initializing data. An alternative way to initialize the growth rates 'r', the initial population sizes 'c' and to (optionally) set the parameters of the proposal distributions of the MCMC. If not specified growth rates are by default all set to 1 and initial population sizes correspond to the first observed read count. This option can for example be used to continue an MCMC run that has not been run long enough from the previous accepted sample. Note that in this case though, the burn-in has to be set to 0.
-hours		If the -hours option is passed time points are assumed to be measured in hours. By default time is assumed to be measured in generations.

Output		
-logLTS		If the -logLTS option is passed a time series of log-likelihoods is written to file.
-ESS		If the -ESS option is passed the effective sample size (ESS) is calculated every 1000 accepted samples and written to file.
-screen		If the -screen option is passed additional output will be written to screen. This option is mainly for inspection purposes.
Random numbers		
-seed	integer, $ > 0	Sets the random number seed. By default the random number seed is created automatically based on computer run time. This option is mainly for inspection purposes.

4 MCMC output

By default, the MCMC program outputs the raw growth rate and initial population sizes samples, along with two separate files containing summary and diagnostic statistics. The same output can optionally be generated for the log-likelihood data. Each file starts with a list of the input/parameters. For all files, data is written in tab-separated format, such that it be displayed nicely and easily with any spreadsheet application (such as Excel or OpenOffice).

Table 3 – A summary of the output of the MCMC program.

File	Parameter	Description
.*_R		
	sample	Consecutive number of samples. Sample '0' gives the initial values.
	r.*	Sampled value for the growth rate 'r' for all mutants.
.*_C		
	sample	Consecutive number of samples. Sample '0' gives the initial values.
	c.*	Sampled value for the initial population size 'c' for all mutants.
.*_logLTS		
	sample	Consecutive number of samples. Sample '0' gives the initial values.
	logL	Log-likelihood for the current sampled values for the initial population size 'c' and the growth rate 'r'.
.*_R_quantiles		
	protID	Protein ID as specified by the input file.
	mutant	Consecutive number of mutant identifier 'r.*'.
	i%	Values for the i%-quantile of all samples of the growth rate 'r' for each mutant, where $i = 0, 1, 2.5, 5, 25, 50, 75, 95, 97.5, 99, 100$.
.*_C_quantiles		
	protID	Protein ID as specified by the input file.
	mutant	Consecutive number of mutant identifier 'c.*'.
	i%	Values for the i%-quantile of all samples of the initial population size 'c' for each mutant, where $i = 0, 1, 2.5, 5, 25, 50, 75, 95, 97.5, 99, 100$.
.*_logL_quantiles		
	logL	Log-Likelihood identifier.
	i%	Values for the i%-quantile of all samples of the log-likelihood, where $i = 0, 1, 2.5, 5, 25, 50, 75, 95, 97.5, 99, 100$.
.*_ess		
	sample	Number of samples after which effective sample size (ESS) is calculated. Note that the ESS is calculated every 'setSize' samples.
	minESS	Minimum effective sample size computed for any parameter of interest (i.e., growth rate 'r', initial population size 'c' and log-likelihood). Note that ESS $\leq$ *sample*.
	r.*	ESS for growth rate 'r.* for all mutant'.
	c.*	ESS for initial population size 'c.*' for all mutants.
	logL	ESS for log-likelihood.
	acceptRatio	Overall acceptance ratio. To ensure high efficiency of the MCMC, the width of the proposal distributions should be chosen such that the acceptance ratio is between $0.15 - 0.45$. Performance is maximal when the acceptance ratio is around 0.25. During the burn-in period the widths of the proposal distributions are automatically tuned to ensure that the acceptance is close to optimal, and efficiency is maximized during sampling. Thus, a sufficiently long burn-in period not only increases the chance that recorded samples are actually taken from the posterior distribution, but also that the width of the proposal distribution is set appropriately.

.*_Diag_R

protID Protein ID as specified by the input file.

mutant.* Consecutive number of mutant identifier 'r.*'.

HD(*) Hellinger distance (HD) between sets of samples from two probability distributions. Note that HD is bounded by $0 \leq \text{HD} \leq 1$ and can be used to inspect the similarity between two distributions, where $\text{HD} = 0$ corresponds to no divergence and $\text{HD} = 1$ corresponds to no common support between the distributions. The HD can be used to diagnose the MCMC in terms of its burn-in and whether samples obtained at different points of time came (most likely) from the same (posterior) distribution. Note that one cannot determine if the MCMC chain has truly converged, but only if a chain is internally similar. Here, the HD is calculated for up to 10 equally sized sets of consecutive samples from the MCMC simulation. To obtain sufficient statistical power, the HD between two sets of samples is calculated only if each set consisted of at least 1000 samples. If the total number of samples is less than $10 \times 1000 = 10000$, the number of batches is chosen such that the total number of samples is divided into sets of samples of size 1000 each. If the total number of samples exceeds 10000, the number of samples per set is given by the total number of samples divided by 10 (i.e., the maximal number of batches). If the HD between sets of samples is less than 0.1 the distribution of posterior samples shows a high degree of similarity; if $0.1 \leq HD \leq 0.3$ the distribution of posterior samples are still quite similar, but may require closer inspection; if $0.3 \leq HD \leq 0.5$ sets of samples are vaguely similar and should be inspected more closely; a $HD > 0.5$ indicates strong dis-similarity between sets of samples and could be an indicator that all samples that were taken before might not be from the posterior distribution and should be discarded as burn-in. Note that the HD depends on the degree of autocorrelation between samples. Thus, a high HD might not necessarily indicate that samples were obtained from different sampling distributions, but poor mixing (i.e., a low ESS) for the parameter of interest. For details see Boone et al. (2014).

mean The mean of the posterior distribution for the parameter of interest.

SD The standard deviation (SD) of the posterior distribution for the parameter of interest calculated with respect to the total number of samples.

median The median of the posterior distribution for the parameter of interest.

2.5% The 2.5% quantile of the posterior distribution for the parameter of interest.

97.5% The 97.5% quantile of the posterior distribution for the parameter of interest.

ESS The effective sample size for the parameter of interest.

minHD The minimum HD calculated between consecutive batches. If there are not enough samples (more than 2000) to calculate the HD this field will read -1.

maxHD The minimum HD calculated between consecutive batches. If there are not enough samples (more than 2000) to calculate the HD this field will read -1.

.*_Diag_C

protID Protein ID as specified by the input file.

mutant Consecutive number of mutant identifier 'c.*'.

HD(*) Hellinger distance (HD) between sets of samples from two probability distributions. Note that HD is bounded by $0 \leq \mathrm{HD} \leq 1$ and can be used to inspect the similarity between two distributions, where $\mathrm{HD} = 0$ corresponds to no divergence and $\mathrm{HD} = 1$ corresponds to no common support between the distributions. The HD can be used to diagnose the MCMC in terms of its burn-in and whether samples obtained at different points of time came (most likely) from the same (posterior) distribution. Note that one cannot determine if the MCMC chain has truly converged, but only if a chain is internally similar. Here, the HD is calculated for up to 10 equally sized sets of consecutive samples from the MCMC simulation. To obtain sufficient statistical power, the HD between two sets of samples is calculated only if each set consisted of at least 1000 samples. If the total number of samples is less than $10 \times 1000 = 10000$, the number of batches is chosen such that the total number of samples is divided into sets of samples of size 1000 each. If the total number of samples exceeds 10000, the number of samples per set is given by the total number of samples divided by 10 (i.e., the maximal number of batches). If the HD between sets of samples is less than 0.1 the distribution of posterior samples shows a high degree of similarity; if $0.1 \leq HD \leq 0.3$ the distribution of posterior samples are still quite similar, but may require closer inspection; if $0.3 \leq HD \leq 0.5$ sets of samples are vaguely similar and should be inspected more closely; a $HD > 0.5$ indicates strong dis-similarity between sets of samples and could be an indicator that all samples that were taken before might not be from the posterior distribution and should be discarded as burn-in. Note that the HD depends on the degree of autocorrelation between samples. Thus, a high HD might not necessarily indicate that samples were obtained from different sampling distributions, but poor mixing (i.e., a low ESS) for the parameter of interest. For details see Boone et al. (2014).

mean The mean of the posterior distribution for the parameter of interest.

SD The standard deviation (SD) of the posterior distribution for the parameter of interest calculated with respect to the total number of samples.

median The median of the posterior distribution for the parameter of interest.

2.5% The 2.5% quantile of the posterior distribution for the parameter of interest.

97.5% The 97.5% quantile of the posterior distribution for the parameter of interest.

ESS The effective sample size for the parameter of interest.

minHD The minimum HD calculated between consecutive batches. If there are not enough samples (more than 2000) to calculate the HD this field will read -1.

maxHD The minimum HD calculated between consecutive batches. If there are not enough samples (more than 2000) to calculate the HD this field will read -1.

.*_Diag_logL

logL Log-likelihood tag.

mutant Consecutive number of mutant identifier 'c.*'.

HD(*) Hellinger distance (HD) between sets of samples from two probability distributions. Note that HD is bounded by $0 \leq \text{HD} \leq 1$ and can be used to inspect the similarity between two distributions, where HD $= 0$ corresponds to no divergence and HD $= 1$ corresponds to no common support between the distributions. The HD can be used to diagnose the MCMC in terms of its burn-in and whether samples obtained at different points of time came (most likely) from the same (posterior) distribution. Note that one cannot determine if the MCMC chain has truly converged, but only if a chain is internally similar. Here, the HD is calculated for up to 10 equally sized sets of consecutive samples from the MCMC simulation. To obtain sufficient statistical power, the HD between two sets of samples is calculated only if each set consisted of at least 1000 samples. If the total number of samples is less than $10 \times 1000 = 10000$, the number of batches is chosen such that the total number of samples is divided into sets of samples of size 1000 each. If the total number of samples exceeds 10000, the number of samples per set is given by the total number of samples divided by 10 (i.e., the maximal number of batches). If the HD between sets of samples is less than 0.1 the distribution of posterior samples shows a high degree of similarity; if $0.1 \leq HD \leq 0.3$ the distribution of posterior samples are still quite similar, but may require closer inspection; if $0.3 \leq HD \leq 0.5$ sets of samples are vaguely similar and should be inspected more closely; a $HD > 0.5$ indicates strong dis-similarity between sets of samples and could be an indicator that all samples that were taken before might not be from the posterior distribution and should be discarded as burn-in. Note that the HD depends on the degree of autocorrelation between samples. Thus, a high HD might not necessarily indicate that samples were obtained from different sampling distributions, but poor mixing (i.e., a low ESS) for the parameter of interest. For details see Boone et al. (2014).

mean The mean of the posterior distribution for the parameter of interest.

SD The standard deviation (SD) of the posterior distribution for the parameter of interest calculated with respect to the total number of samples.

median The median of the posterior distribution for the parameter of interest.

2.5% The 2.5% quantile of the posterior distribution for the parameter of interest.

97.5% The 97.5% quantile of the posterior distribution for the parameter of interest.

ESS The effective sample size for the parameter of interest.

minHD The minimum HD calculated between consecutive batches. If there are not enough samples (more than 2000) to calculate the HD this field will read −1.

maxHD The minimum HD calculated between consecutive batches. If there are not enough samples (more than 2000) to calculate the HD this field will read −1.

.*_Diag_summary

samples Absolute number of accepted samples taken during the MCMC run.

minESS(c) The minimum effective sample size (ESS) that was observed for any initial population size 'c.*'.

maxACT(c) The maximal auto-correlation time (ACT) that was observed for any initial population size 'c.*'.

minESS(r) The minimum effective sample size (ESS) that was observed for any growth rate 'r.*'.

maxACT(r) The maximal auto-correlation time (ACT) that was observed for any growth rate 'r.*'.

minESS(logL) The minimum effective sample size (ESS) that was observed for the log-likelihood.

maxACT(r) The maximal auto-correlation time (ACT) that was observed for the log-likelihood.

minESS(all) The minimum effective sample size (ESS) that was observed for all parameters.

maxACT(all) The maximal auto-correlation time (ACT) that was observed for all parameters.

acceptRatio
The overall acceptance ratio of accepted (and recorded) samples. To ensure high efficiency of the MCMC, the width of the proposal distributions should be chosen such that the acceptance ratio is between 0.15 − 0.45. Performance is maximal with an acceptance ratio around 0.25. During the burn-in period the width of the proposal distributions is automatically tuned – based on the acceptance ratio – such that the acceptance ratio, when recording samples, has close to maximal efficiency. Thus, a sufficiently long burn-in period not only increases the chance that recorded samples are actually taken from the posterior distribution, but also that the width of the proposal distribution is set appropriately.

growthRateSD
Standard deviation of the proposal distribution of growth rates 'r' drawn from a Gaussian distribution after auto-tuning.

popSizeSD
Scale parameter of the proposal distribution of initial population sizes 'c' drawn from a Cauchy distribution after auto-tuning.

.*_initialRC

This file prints the last accepted values of the MCMC run so that these could be used as initial values, e.g., to continue an MCMC run that has not yielded enough independent samples. The first line gives the last sampled growth rates 'r.*', the second line gives the last sampled initial population sizes 'c.*', and the third line gives the standard deviation of the proposal distribution of growth rates 'r' (Normal distribution) and the scale parameter of the proposal distribution of initial population sizes 'c' (Cauchy distribution) after auto-tuning.

4.1 Combining Files

In case the data set has been split into multiple subsets to enhance computational performance (see '-g' option in Tab. 1), we provide scripts to assemble the individual MCMC output files for each sub-data set into a single file (restricted to 'diagnostic', 'quantiles', and 'initial population size' and 'growth rate' files). `Combine_All.sh` is a wrapper which executes all 'Combine' scripts (i.e., `Combine_Diagnostic_C.sh`, `Combine_Diagnostic_R.sh`, `Combine_Quantiles_C.sh`, `Combine_Quantiles_R.sh`, `Combine_PopSizes_C.sh` and `Combine_GrowthRates_R.sh`). It furthermore removes the time stamp from all `empiricIST_MCMC` output files and (optionally) deletes the files that will not be combined. More information on all the script files is given in Table 4.

Please ensure that the scripts have the appropriate file permissions to perform the operations. File permissions can be adjusted by using the command-line and typing

```
chmod 755 <file> .
```

Table 4 – A summary of the 'combine files scripts'.

Script	No. of passed variables	Description
`Combine_All.sh`	6	This script removes the time stamp from all MCMC output files, executes all individual combine scripts, and (optionally) deletes the files that will not be combined. To execute the script type ./`Combine_All.sh` pathToPerlRename pathToData prefixFileName suffixFileName maxIndex deleteFiles . pathToPerlRename: Provide full path to 'rename.pl'. pathToData: Provide full path to data folder. prefixFileName: Provide file name prefix. suffixFileName: Provide file name suffix. maxIndex: The number of sub-data sets the original data has been split-up to. deleteFiles: When set to '1' files that will not be combined will be deleted. Set to 0 otherwise. For an example see Figure 3. Information on the individual combine scripts can be found below.
`Combine_Diagnostic_C.sh`	4	This script combines all MCMC ouput files of type '`Diagnostic_C` of all sub-data sets into a single file. Note that the reference sequence will be deleted, since it does not contain any relevant information. To execute the script type ./`Combine_Diagnostic_C.sh` pathToData prefixFileName suffixFileName maxIndex . pathToData: Provide full path to data folder. prefixFileName: Provide file name prefix. suffixFileName: Provide file name suffix. maxIndex: The number of sub-data sets the original data has been split-up to.
`Combine_Diagnostic_R.sh`	4	As above, but for growth rates 'R'.
`Combine_Quantiles_C.sh`	4	This script combines all MCMC ouput files of type '`Quantiles_C` of all sub-data sets into a single file. Note that the reference sequence will be deleted, since it does not contain any relevant information. To execute the script type ./`Combine_Quantiles_C.sh` pathToData prefixFileName suffixFileName maxIndex . pathToData: Provide full path to data folder. prefixFileName: Provide file name prefix. suffixFileName: Provide file name suffix. maxIndex: The number of sub-data sets the original data has been split-up to.
`Combine_Quantiles_R.sh`	4	As above, but for growth rates 'R'.

`Combine_PopSizes_C.sh`	4	This script combines all MCMC ouput files of type '`_C` of all sub-data sets into a single file. Note that the reference sequence will be deleted, since it does not contain any relevant information. To execute the script type ./`Combine_PopSizes_C.sh` pathToData prefixFileName suffixFileName maxIndex . pathToData: Provide full path to data folder. prefixFileName: Provide file name prefix. suffixFileName: Provide file name suffix. maxIndex: The number of sub-data sets the original data has been split-up to.
`Combine_GrowthRates_R.sh`	4	As above, but for growth rates 'R'.
`RemoveMCMCTimeStamp.sh`	3	This script removes the time stamp from all `empiricIST_MCMC` output files. To execute the script type ./`RemoveMCMCTimeStamp.sh` pathToPerlRename pathToData prefixFileName . pathToPerlRename: Provide full path to 'rename.pl'. pathToData: Provide full path to data folder. prefixFileName: Provide file name prefix.

4.2 Concatenating Files

When an MCMC analysis has been continued across multiple runs (i.e., by starting another MCMC run from the last accepted sample; see -initial option above), the 'ConcantenateData.sh' script can be used to produce a single file of all sampled growth rates 'r', initial population sizes 'c' and log-likelihoods, respectively.

$$\underbrace{\text{MCMCInput}}_{\text{prefixName}} - \overbrace{99}^{\text{maxIndex}} - \underbrace{\text{HGW\_VIE}}_{\text{suffixName}}\overbrace{\text{\_08\_06\_2015\_142000\_}}^{\text{date stamp}}\underbrace{\text{R}}_{\text{type}}.\mathit{txt}$$

Figure 3 – Illustration of how to set the command line arguments for the combine scripts.

Table 5 – A summary of the 'combine files scripts'.

Script	No. of passed variables	Description
`ConcatenateData.sh`	8	This script removes the time stamp from all MCMC output files, (optionally) deletes the files that will not be combined, and concatenates the individual files. To execute the script type ./`ConcantenateData.sh` pathToData pathToMove pathToPerlRename prefixFileName suffixFileName simIdentifier maxIndex deleteFiles . pathToData: Provide full path to data folder. pathToMove: Provide full path where to move data. pathToPerlRename: Provide full path to 'rename.pl'. prefixFileName: Provide file name prefix. suffixFileName: Provide file name suffix. simdentifier: Provide indicator for which data to concatenate: R: growth rates, C: initial population sizes, logLTS: log-likelihood maxIndex: The number of MCMC runs the original data has been split-up to. For an example see Figure 3. deleteFiles: When set to '1' files that will not be combined will be deleted. Set to 0 otherwise.

Once all samples have been combined into a single file, statistics and diagnostics can be calculated by using the provided python programs `empiricIST_MCMC_Statistics.py` and `empiricIST_MCMC_DiagSummary.py`, respectively. Taking the raw samples as input, `empiricIST_MCMC_Statistics.py` creates the `quantiles`- and the `Diag` files as if obtained from a single MCMC run (for detail see above). The python program `empiricIST_MCMC_DiagSummary.py` can then be used to create the `Diag_summary` file over all individual diagnostic statistics (see above for details). Note though, that unlike the file produced from a single MCMC run, the one created by the python program does not report the standard deviation of the proposal distribution of growth rates 'r' and initial population sizes 'c', nor the acceptance ratio. Furthermore, it can only be used when the individual log-likelihood, growth rate and initial population size files exist. Table 6 and table 7 give an overview of the options for each program.

Table 6 – A summary of the options of the `empiricIST_MCMC_Statistics.py` program.

Short/Long option	Accepted values	Description
-h, –help	none	When the '-h'-option is invoked, a short documentation on the usage of the program is shown. Note that, if this option is invoked, the python program is not executed.
-f, –file=	string	The '-f'-option is a mandatory option, which passes the name of the data input file to the python program. Files created by the python program will take the name of the input file and insert `_Quantiles.txt` and `_Diag_`, respectively.

Table 7 – A summary of the options of the `empiricIST_MCMC_DiagSummary.py` program.

Short/Long option	Accepted values	Description
-h, –help	none	When the '-h'-option is invoked, a short documentation on the usage of the program is shown. Note that, if this option is invoked, the python program is not executed.
-f, –file=	string	The '-f'-option is a mandatory option, which passes the base name of the data input files to the python program. For example, if the file names are `DFE_Diag_logTS.txt`, `DFE_Diag_C.txt` and `DFE_Diag_R.txt` the base name to be passed is `DFE_Diag`. The output file created by the python program will take the name of the input file and prepend `_summary.txt`.
-s, –samples=	integer	The '-s'-option is a mandatory option, which passes the number of (posterior) samples to the python program.

4.3 Visualization of trace data

Additionally, there is a shell script 'FormatTracer.sh' that formats the posterior sample output file (e.g., containing the initial population size 'c' or the growth rate 'r') such that it can be read and analyzed by *Tracer* (Rambaut et al. 2014). *Tracer* is a graphical tool for visualization and diagnostics of MCMC output that, for instance, displays the posterior distribution and its credibility interval, calculates the effective sample size (ESS; note that values might be slightly different from those calculated by the `empiricIST_MCMC` program since we use a more accurate but computational more intensive algorithm), and shows the trace of the posterior samples. Note that the input file is excepted to be formatted as the output file created from the `Combine_PopSizes_C.sh`/`Combine_GrowthRates_R.sh` script. Otherwise the provided `Create_TailShapeFileR.sh` script can be used to obtain a correctly formatted input file (please see below for details).

./`FormatTracer.sh` pathToData fileName .

pathToData: Provide full path to data folder.

fileName: Provide file name without file extension (e.g., '.txt').

5 DFE tail shape estimation

The growth rate posterior samples obtained from the `empiricIST_MCMC` program can be used to estimate the shape of the beneficial tail of the distribution of fitness effects (DFE). As part of the *empiricIST* software package, we provide a python script – empiricIST_MCMC_TailShape.py that fits a generalized pareto distribution to the observed beneficial mutations by maximizing the log-likelihood with respect to the shape and scale parameter κ and ψ, respectively. Based on the shape parameter κ one can discriminate between three different domains of attraction – the Weibull, Gumbel and Fréchet domain – each corresponding to a different extreme value distribution. In biological terms, these different domains quantify the level of adaptedness of the organism in its (experimental) environment. In particular, the Gumbel domain ($\kappa = 0$; null model corresponding to 'normal' level of adaptedness) is characterized by an exponential tail, whereas the Weibull domain ($\kappa < 0$; better adapted) has lighter than exponential tails, and the Fréchet domain ($\kappa > 0$; less well adapted) has heavier than exponential tails. For more details please consult Beisel et al. (2007).

While the python script primarily estimates the DFE tail shape parameter κ, there are some additional options that will be detailed here.

The general usage is as follows: The program can be executed by typing

```
python empiricIST_MCMC_TailShape.py [options] .
```

Without specifying any options the program will exit with an error and provide a short documentation on its usage, as it requires the name of the data input file (by invoking the '-f' option). Note that the input file needs to be formatted in a specific way. If the data set had been split into multiple subsets (see '-g' option in Tab. 1), and has been re-assembled by the using the provided shell scripts (see 'Combining Files'), the input file is already correctly formatted and there is nothing that needs to be done. However, if the data set has been analyzed as a whole (i.e., without being split into multiple subsets) the file containing the posterior growth rate samples needs to be reformatted to match the required input format. This can be done by using he provided shell script Create_TailShapeFileR.sh, which creates the input file for the empricIST_MCMC_TailShape.py program and can be executed by typing

./Create_TailShapeFileR.sh pathToData fileName .

pathToData: Provide full path to data folder.

fileName: Provide file name without file extension (e.g., '.txt').

Please ensure that the script has the appropriate file permissions to perform the operations. File permissions can be adjusted by using the command-line and typing

```
chmod 755 Create_TailShapeFile_R.sh .
```

All options and their usage of the `empricIST_MCMC_TailShape.py` program are given in Table 8.

Table 8 – A summary of the options of the `empiricIST_MCMC_TailShape.py` program.

Short/Long option	**Accepted values**	**Description**
-h, –help	none	When the '-h'-option is invoked, a short documentation on the usage of the program is shown. Note that, if this option is invoked, the python program is not executed.
-f, –file=	string	The '-f'-option is a mandatory option, which passes the name of the data input file to the python program. Files created by the python program will take the name of the input file and add option-dependent specific file identifiers. Note that even if a random data set is created (see 'r'-option), a file name must be provided since it serves as the prefix for the output file name.
-m, –missing	none	When the '-m'-option is invoked, the distribution of measured fitnesses is shifted relative to the smallest observed selection coefficient to account for missing data (i.e., selection coefficients too small to have been observed, though this might not be a problem with EMPIRIC data).
-s, –samples=	integer	When the '-s'-option is invoked, the python program will only consider samples with more than 'samples' beneficial mutations for maximum likelihood estimation. Note that when this option is not specified, the default is set to 10.
-l, –lhoodrt	none	When the '-l'-option is invoked, a likelihood-ratio test with null hypotheses $\kappa = 0$ against median($\hat{\kappa}$) is performed. Note that the distribution of the test statistic is generated by using a parametric bootstrap approach with 10.000 samples (Beisel et al. 2007, see).
-r, –random=	double	When the '-r'-option is invoked, 100 random data sets are created with 100 samples each drawn from a generalized pareto distribution with scale parameter $\psi = 1$ and κ passed as command line argument.

By default the program will always create two output files that contain the κ and ψ estimates called `<InputFileName>_TailShape_KappaShape.txt` and `<InputFileName>_TailShape_PsiShape.txt`, respectively. Note that even if a random data set is created (see 'r'-option), a file name must be provided since it serves as the prefix for the output file name. When performing a likelihood-ratio test (by invoking the 'l'-option) an additional file called `<InputFileName>_LRT.txt` is created containing the $\hat{\kappa}$ against which $H_0 : \kappa = 0$ is evaluated. Furthermore, the output file gives the maximum-likelihood estimate $\hat{\psi}_{\kappa 0}$, i.e., the estimated scale parameter ψ restricting $\kappa = 0$, the value of the test statistic

$$-2\log(\Lambda) = 2(\mathcal{L}(\mathbf{X} \mid \hat{\kappa}, \hat{\psi}) - \mathcal{L}(\mathbf{X} \mid 0, \hat{\psi}_{\kappa 0})), \tag{1}$$

where $\mathbf{X}$ denotes a single vector of posterior samples of growth rates and $\mathcal{L}$ is the log-likelihood function, and the associated p-value along with the sample size.

Note that power critically depends on sample size as has been discussed in Beisel et al. (2007) and del Castillo and Serra (2015) (the latter rather in the context of estimating κ accurately).

References

Bank, C., R. T. Hietpas, A. Wong, D. N. Bolon, and J. D. Jensen, 2014. A bayesian mcmc approach to assess the complete distribution of fitness effects of new mutations: Uncovering the potential for adaptive walks in challenging environments. Genetics 196:841–852.

Beisel, C. J., D. R. Rokyta, H. A. Wichman, and P. Joyce, 2007. Testing the extreme value domain of attraction for distributions of beneficial fitness effects. Genetics 176:2441–2449.

Boone, E. L., J. R. Merrick, and M. J. Krachey, 2014. A hellinger distance approach to mcmc diagnostics. Journal of Statistical Computation and Simulation 84:833–849.

del Castillo, J. and I. Serra, 2015. Likelihood inference for generalized pareto distribution. Computational Statistics And Data Analysis 83:116 – 128.

Rambaut, A., M. Suchard, D. Xie, and A. Drummond, 2014. Tracer v1.6.