fastGRiP User Manual

Daphne Ezer[1,2,*], Nicolae Radu Zabet[1,2,*,‡] and Boris Adryan[1,2,†]

[1] Cambridge Systems Biology Centre, University of Cambridge, Tennis Court Road, Cambridge CB2 1QR, UK;

[2] Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK

[*] D.E. and N.R.Z. contributed equally to this work.

[†]Corresponding author: ba255@cam.ac.uk, [‡]Correspondence can also be addressed to: n.r.zabet@gen.cam.ac.uk

Introduction to fastGRiP

FastGRiP is a web service to simulate the process by which TFs search for their binding sites along the DNA by facilitated diffusion.

It can be accessed by going to the website: $www.fastgrip.edu$. The paper describing fastGRiP is available at: TBD. FastGRiP is a semi-analytic approximation of GRiP, which is available at: $http://logic.sysbiol.cam.ac.uk/grip/$ and described in Zabet and Adryan (2012b).

Input

The top navigation bar on the website allows the user to toggle between different modes by which they can run fastGRiP. Each of these modes uses the same core algorithm, but is modified to accept differing input parameters. These modes are: uniform landscape with tau_0 as input, uniform landscape with PWM as input, and non-uniform landscape.

Uniform Landscape: τ_0 as input

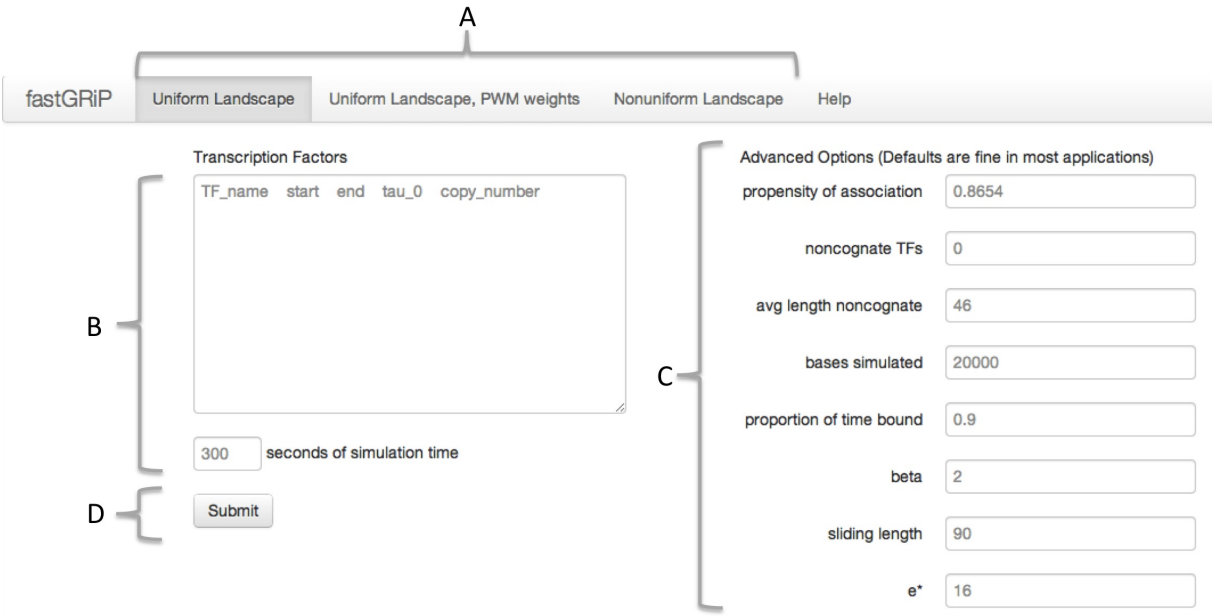

Figure 1: *Uniform landscape screen capture.* This is the first screen that appears when you go to the fastGRiP website. The top navigation bar (A) provides you with a choice of input options. The inputs to the "Uniform Landscape" tab is selected by default. You should select this tab if you know the locations of the TF, their relative *stickiness* to the DNA (t_0) and their abundance/copy number. These should be inputed, as a table, in (B), along with the amount of time you wish to run the simulation. This tab is particularly useful if you wish to study the theoretical properties of a promoter architecture, and care more about the relative binding affinities rather than their absolute binding affinities. One can specify advanced options in (C), as described below. When the data is filled in, the form should be submitted (D).

...

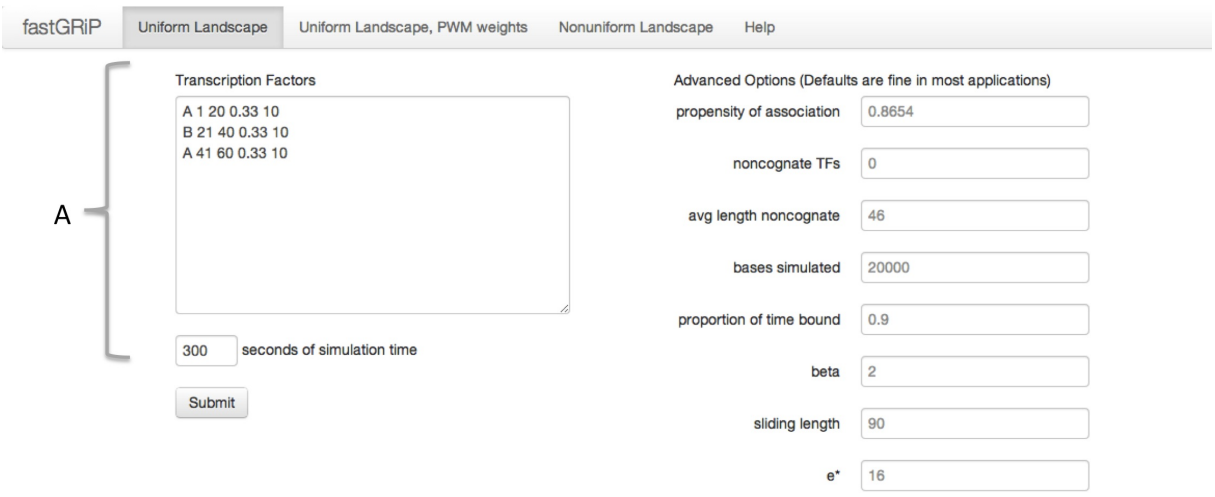

Figure 2: *Example of Input to Uniform Landscape.* (A) This is an example of a properly filled out form. In specific, this form shows how to input an *ABA configuration* with 0 bp separation, run for 300 seconds.

Parameters

transcription factors

This box requires a tabular input of TF binding site by TF name, start location, end location, τ_0, and TF abundance. Include a description of 1 TF binding site per line. Any whitespace can be used as a delimiter. TF binding sites that are associated with the same TF name can be bound by the same TF. Be sure that TFs with the same name have the same abundance.

Example: ABA configuration

A 1 10 3.3 100
B 11 20 0.33 10
A 21 30 3.3 100

seconds of simulation time

The length of biological time to simulate. We recommend originally running the simulation for 100 seconds, and then re-running simulation for longer times if the application requires it. Note that *E.coli* has a cell cycle of 3000 seconds.

propensity of association

The propensity of association is the rate at which TFs bind to the DNA by 3D diffusion. This is the variable called k_{assoc} in Zabet and Adryan (2012a). Note that the *propensity of association* and the *number of base pairs* (described below) must be adjusted together as described in Zabet (2012). The default value has been adjusted to match 20000 bp being simulated.

noncognate TFs

The number of non cognate TFs is the number of nonspecifically bound proteins that are present around the promoter. When there are more non cognate molecules, the rate of binding is decreased, because the ability to undergo 1D diffusion is reduced and because the binding site might be temporarily bound by a non cognate protein.

avg length noncognate

The average length of the non cognate is the number of base pairs that a non cognate protein covers when it is randomly bound to a piece of DNA. It is by default $46bp$ because this was the value selected in Zabet and Adryan (2012a), because there are about 4.6 million base pairs in the *E.coli* genome and that is divisible by 46. Therefore, the default comes more from legacy than for any biological motivation.

bases simulated

This number represents the number of base pairs that are simulated (length of DNA). In reality, fastGRiP does not explicitly depend on the length of the DNA sequence, so this number is used to reverse the adjustment of *the propensity of association* as described in Zabet (2012). We allow users to input *the number of base pairs* and *the propensity of association*, adjusted, instead of just the raw propensity of association, so that they can easily input the published values from GRiP.

proportion of time bound

The proportion of the time bound is the percentage of the time a TF is undergoing a random walk along the DNA rather than conducting 3D diffusion.

beta

Beta (β) is a parameter that adjusts the exponent of binding affinity, as described in Zabet and Adryan (2012a).

sliding length

The sliding length is the average distance that a TF slides during its random walk along the DNA.

e-star: e^*

e^* is a parameter describing the energy of a binding site, as described in Zabet and Adryan (2012a).

Uniform Landscape: PWM score as input

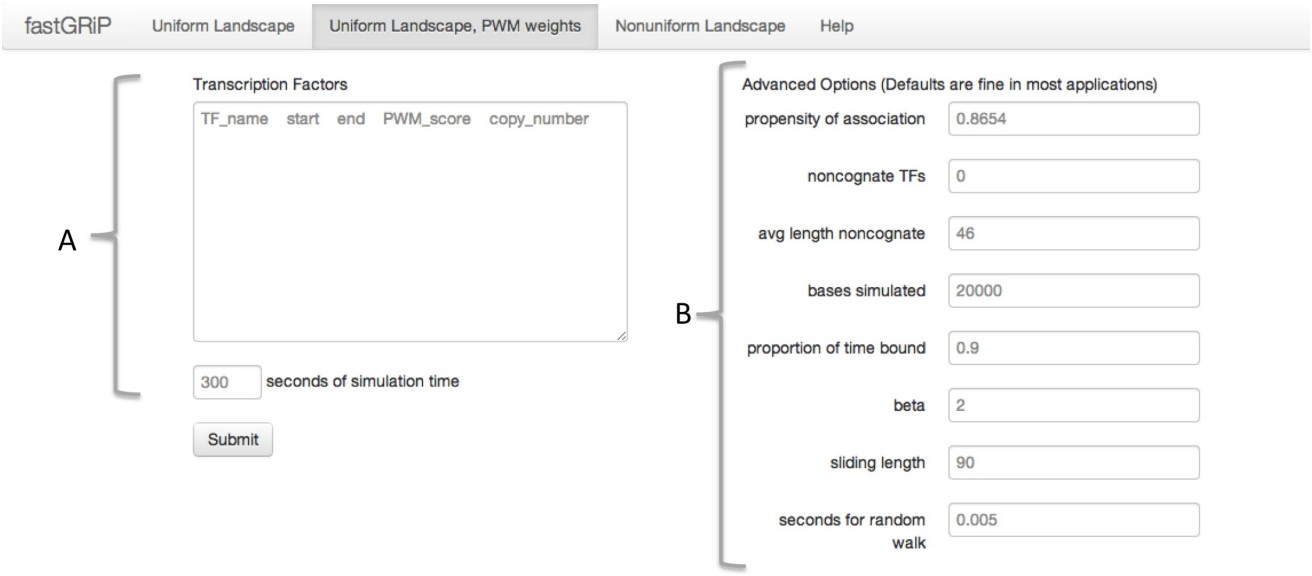

Figure 3: *Uniform Landscape, PWM weights.* In the *Uniform Landscape, PWM weights* tab, you can include PWM scores to correlate with binding affinity, rather than the arbitrary parameter t_0, as in the Uniform Landscape case, as seen in (A). In the advanced parameter list (B), we see that most of the parameters are the same as before, except that instead of adjusting e^*, you can adjust the amount of time a TF spends during its random walk along the DNA.

Parameters

Many of the parameters are the same as in the case of *Uniform Landscape:* τ_0. Here we highlight the differences.

transcription factors
This box requires a tabular input of TF binding site by TF name, start location, end location, *PWM score*, and TF abundance.

Example: ABA configuration

A 1 10 12.3 100
B 11 20 9 10
A 21 30 13.2 100

seconds for random walk
The seconds for a random walk is the average amount of time a TF spends during its random walk along the DNA to a region that does not contain its binding site.

Parameters

Many of the parameters are the same as in the case of *Uniform Landscape: PWM score as input.* Here we highlight the differences.

Transcription Factor Info
This box requires a tabular input of TF binding site by TF name, start location, end location, and TF abundance. Note that TF names must be those with PWM scores in RegulonDB, specifically

Nonuniform Landscape

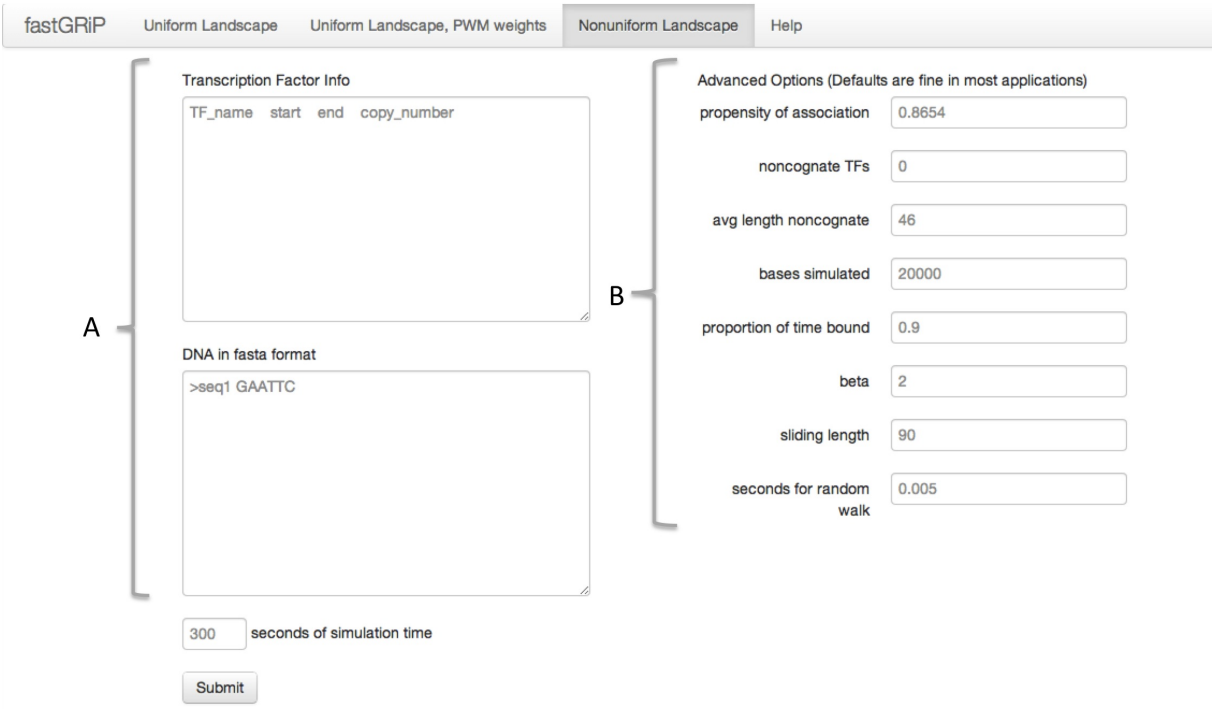

Figure 4: *Nonuniform Landscape*. If you have evidence to suggest that properties of the local affinity landscape would influence binding, then use this tab. In (A), input the TF name, the location, and the abundance/copy number and the affinity landscape will be calculated for you by the PWMs. In this tab, we ONLY allow you to use TF names of TFs that are found in RegulonDB. This is because calculating one of the parameters (t_0) from a new PWM is time intensive, although we may add that functionality in a later version of this website. Afterwards, input the DNA sequence in fasta format. Note that the advanced parameter list (B) is the same as in the case of Uniform Landscape, PWM Weights.

http://regulondb.ccg.unam.mx/menu/download/datasets/files/PSSMSet.txt. Also note that there is no need to specify binding affinity by any metric, because these will be calculated based on the DNA sequence.

Example:

AgaR 1001 1010 100
HNS 1011 1020 10
Lrp 1021 1030 100

DNA in fasta format

This is the DNA sequence that will be analyzed, in fast format. This DNA sequence will be used to calculate PWM scores and will be used to calculate the specific time spent during the random walk around the binding site, based on the local affinity landscape. It must include at least as many bp on either side of each binding site as the sliding length.

Output

There are two graphs that are produced by fastGRiP: 1) a network representation of the rate of state transitions at equilibrium and 2) an interactive scatterplot of the frequency of being in a configuration, over time. There are also a number of tabs that allow the user to access the raw output data: the network data, the time course data, the first occupancy (TF) data, and the first occupancy (config) data.

GUI

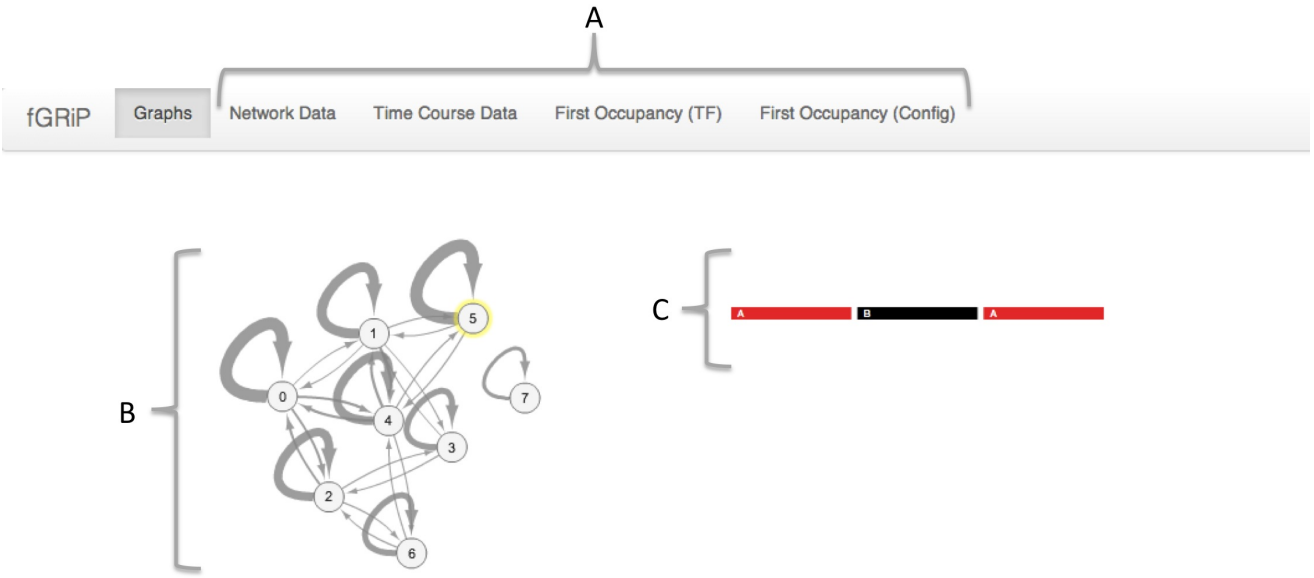

Figure 5: *fastGRiP Output: Network Depiction.* This is the screen that appears when the simulation is complete. The navigation bar (A) provides access to the raw output data of the simulation. The data formats of these will be described later. In (B), there is a network displayed in which each circle represents a promoter configuration, numbered in binary. For instance, in state 0, none of the TFs are bound, in state 1, the first TF is bound, etc. The thickness of the transitions represents the number of transitions between states that occur across 5 seconds of simulation time. To see what configuration each state corresponds to, click on a node: it will be highlighted and a drawing of the configuration will appear in (C). In (C), each box represents a binding site and the red boxes are those that are bound by a TF.

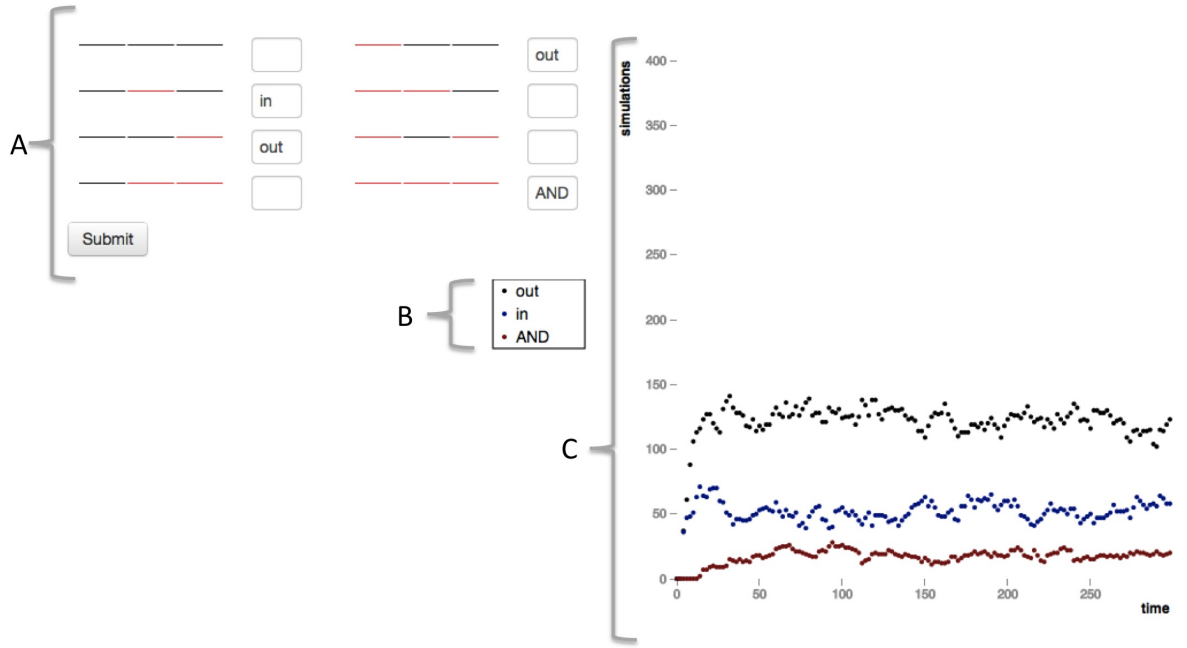

Figure 6: *fastGRiP Output: Interactive Scatterplot.* (A) lists all of the promoter configurations, in which a red line indicates that a TF is bound at that position. You can add labels to the boxes to the right of each configuration and click submit. (B) shows the legend, which lists the labels with their corresponding colors and (C) shows the scatterplot, which graphs the number of simulations that have a particular configuration over time. Note that if multiple configurations are labelled the same, the graph with display the sum of the number of simulations in that configuration at a point in time.

Raw Data

Network Data

The network data is in a cytoscape-readable JSON format. First there is a data schema that describes the structure of the graph, then there is a list of edges, and finally a list of nodes. The nodes are enumerated in the same way as in the graphical display described in the GUI section. The edges each have a weight value associated with them: the way this weight is calculated is by taking the last half of the simulation time and counting the number of times a particular transition happened across each five second interval. This is not the same as the Markov Chain; for instance, let's say that state 1 always transitions to state 2, which is an unstable state that quickly transitions to state 3. In the Markov Chain, you would see $1- > 2- > 3$, but in this network diagram you would see $1- > 3$, because you rarely observe the system in state 2. Therefore, this network tells you more about the logic of what goes on in the system, since it incorporates both the Markov Chain topology and information about the transition times and state stabilities.

Time Course Data

The time course data is the raw data used in the interactive scatterplots. Each line represents a configuration, ordered by the same numbering system used in the GUI. The numbers, separated by spaces, represent the number of simulations (out of 400) that were in that configuration at a given time point; the time points are sampled every 5 seconds.

First Occupancy (TF)

The first occupancy (TF) data represents the arrival times of each TF in each simulation. Each column represents a different TF, ordered by the start positions of the TF. Each row represents a different simulation (there are 400 simulations run). The data represents the first time the TF reaches its binding site in a particular simulation.

First Occupancy (Config)

The first occupancy (config) is the same as first occupancy (TF), except that it provides the arrival times for each configuration, not just each TF. The configurations are ordered as they are in the GUI.

References

Zabet, N. R. (2012). System size reduction in stochastic simulations of the facilitated diffusion mechanism. *BMC Systems Biology*, 6(1):121.

Zabet, N. R. and Adryan, B. (2012a). A comprehensive computational model of facilitated diffusion in prokaryotes. *Bioinformatics*, 28(11):1517–1524.

Zabet, N. R. and Adryan, B. (2012b). GRiP: a computational tool to simulate transcription factor binding in prokaryotes. *Bioinformatics*, 28(9):1287–1289.