Journal of
Computer Science & Systems Biology

Cserháti, J Comput Sci Syst Biol 2012, 5:2
DOI: 10.4172/jcsb.1000093

Research Article Open Access

Prommute – A Promoter Mutation Simulation for Modeling the Evolution of Genetic Regulatory Elements

Mátyás Cserháti*

Institute of Chemistry, Eötvös University (ELTE), 1117 Budapest, Pázmány Péter sétány 1/A, Hungary

Abstract

Studying mutations in promoter sequences has revolutionized molecular genetics by taking into account changes in the sequence which facilitate functional changes. Our knowledge of such changes can be furthered by tracking these changes as they occur after each base pair mutation. Although these mutations cannot be repeated or directly observed throughout molecular evolution, they can be modeled to give a feel of the dynamics of how regulatory elements are formed through time.

This article presents PromMute, the graphical promoter mutation simulation, designed to model the appearance of transcription factor binding sites in promoters through single base pair mutations. It is capable of tracking the formation of a number of transcription factor binding sites, either from yeast, or supplied by the user through a number of generations applying natural selection. The program is compared to existing programs such as ev and PPE (Probability of Promoter Evolution).

Different kinds of sample test simulations were done with the program, including studying the number of generations needed for the appearance of a given motif due to random mutations as well as the dynamics of motif turnover.

PromMute is capable of modelling the transcription factor binding sites and scoring them more realistically. The sample test cases presented in the article show that longer transcription factor binding sites take a longer time to form, and that such motifs are also more prone to deformation by back mutations. The program is also useful for researchers who wish to study motif turnover of their own specified motifs.

Keywords: Promoter, Mutation, Simulation, Evolution, Yeast, Motif, TFBS

Abbreviations: ABF1: Autonomously replicating sequence Binding Factor 1; CSRE: Carbon Source-Responsive Element; GAL4: GALactose metabolism; GCR1: GlyColysis ReguResistance to Lethality of lation 1; MCB (RPN): Regulatory Particle Non-ATPase; MIG1: Multicopy Inhibitor of GAL gene expression; PPE: Probability of Promoter Evolution; PWM: Position Weight Matrix; RGI: Rate of Generation Increase; RLM1: Resistance to Lethality of MKK1P386 overexpression 1; SCPD: Saccharomyces Cerevisiae Promoter Database; SMP1: Second MEF2-like Protein 1 1; TFBS: Transcription Factor Binding Site

Introduction

In recent decades evolutionary theory has emphasized studying the evolution of genetic regulatory elements. This is because changes in genetic regulatory networks are purported to explain different kinds of morphological and physiological changes in organisms and also lend an explanation as to how genomes evolve [1]. Small changes in regulatory genetic elements are therefore be the cause behind changes in phenotype, thereby affecting evolutionary development [2].

Modelling the evolution of genetic regulatory elements is a relatively new area in molecular genetics. A number of articles discuss different models of transcription factor binding site (TFBS) distribution and turnover from an information theory viewpoint [3-5]. These include finding regulatory motifs based on their information content which mirrors the difference in base content between the motif and the genome as a whole.

The thermodynamic interactions between the surface of a transcription factor and the surface of the DNS determine how strong these two macromolecules bind to each other. The binding rates of these transcription factors to their binding sites is correlated to the information content of the motifs themselves, denoted by R_i [6]. R_i is determined by the matrix $R_{iw}(b,l)$, which is used to determine the information content of a given motif. An R_i value of 0 seperates binding sites (above 0) from non-binding sites (below 0). The majority of binding sites have an information content around $R_{sequence}$, which is the average information content of the motif. For more about the relationship between binding strength and the information content of TFBS's, see the two publications of Schneider [7,8].

The ev program of Schneider [9] is worth mentioning which is based on the groundbreaking work done in actually simulating molecular evolution in promoter sequences by measuring the growth in information content of TFBS's. The ev program simulates the formation of only a single kind of TFBS at 16 different sites within a single promoter. This is however unrealistic, since most TFBS's usually occur only a few times within promoters. Another short-coming of the ev program is that instead of defining the fit between TFBS's and transcription factors in this way, it unrealistically models the fitting of transcription factors to their individual TFBS's using a complicated method involving twos complement matrixes. Another program, PPE of Stone and Wray [10] models the time needed in their own simulation

***Corresponding author:** Mátyás Cserháti, Institute of Chemistry, Eötvös University (ELTE), 1117 Budapest, Pázmány Péter sétány 1/A, Hungary, Tel: +36-1-372-2500/1108; Fax: +36-1-372-2592; E-mail: csmatyi@chem.elte.hu

Received May 24, 2012; **Accepted** July 17, 2012; **Published** July 20, 2012

Citation: Cserháti M (2012) Prommute – A Promoter Mutation Simulation for Modeling the Evolution of Genetic Regulatory Elements. J Comput Sci Syst Biol 5: 074-080. doi:10.4172/jcsb.1000093

Copyright: © 2012 Cserháti M. This is an open-access article distributed under the terms of the Creative Commons Attribution License,which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

for new TFBS's to appear through random mutation in *Drosophila* promoters. This program uses real biological sequences and models mutations through stochastic processes, but does not model selection.

The goal of this article is to present a multi-faceted simulation program which realistically models motif turnover in promoter sequences, called PromMute. Since the myriads of base pair mutations which supposedly happened in the evolutionary past cannot be reproduced site-by-site and generation by generation, the PromMute program has been developed to simulate this type of mutation in a set of randomized promoter sequences over a set number of generations in order to visualize the dynamics of TFBS formation *in silico*.

The program simulates the formation of a number of known TFBS's (also called motifs) through random mutations. Any number of TFBS's can be selected from a list of 24 yeast motifs from the Promoter Database of *Saccharomyces cerevisiae* (SCPD) [11]. Different motifs, belonging to other species defined by a position weight matrix can also be added. TFBS's from widely used databases such as Jaspar and Transfac can also be imported and studied. More than one TFBS can be selected in order to study the development of whole TFBS networks. Therefore different kinds of simulations can be devised to simulate whether a given kind of regulatory network can form through mutation.

Materials and Methods

Architecture

PromMute is a stand-alone Windows program, and was implemented in Visual C# in Microsoft Visual Studio 2008, and requires a .NET framework 3.5 for installation. The position weight matrices can be imported either by directly typing them in the program or by opening a simple text file containing the matrix values. The program can be downloaded at the following website: http://prommute.sourceforge.net.

Calculation of the best score

The user may select a number of motifs to be analyzed during the simulation. The same number of position weight matrixes (PWMs), one for each motif, is then used to score the promoter of each of the organisms for every generation. The best score (BS) for all promoters is given in the following equation:

$$BS = \left[\sum_{i=1}^{n_m} \left[\sum_{j=1}^{l_i} PWM_i(b,j) \right]_{\max} \right]_{\max} \quad (1)$$

Here n_m is equal to the number of motifs selected by the user, l_i is the length of the i^{th} motif, and $[\mathrm{PWM}_i(b,j)]_{\max}$ is the highest position weight matrix score for the i^{th} motif, where b is the j^{th} base along the motif. The PWM for each motif is fitted along the entire promoter to see where it fits best, that is, where it's PWM score is the highest. In other words, the program slides each PWM along each promoter sequence in all organisms, and selects the best fit of all PWMs, and does this for all promoters. The promoters are then scored by adding up the maximum PWM score for all motifs. The promoters are then ranked according to their individual scores. This is crucial because in the selection step of the program, the top percentage of all organisms is then selected whose promoters scored the highest. PromMute also keeps track of the maximum best score value, as well as the minimum best score value over all generations in order to visualize how the fitness value of the organisms changes through time.

Simulation halt

The simulation stops when either the generation number limit has been reached selected to be large enough (in our case 10,000 generations), or when all of the selected motifs have appeared in at least one of the organisms. In other words, if the program reaches the generation limit, this means that the motifs cannot form (assuming that the motifs would not form even if time went on until infinity). However, if in at least one of the organisms the maximum PWM score for all of the motifs is greater than or equal to the threshold limit set by the user, then this means that the motifs have formed, and the program stops (meaning that the evolutionary process has achieved its goal of forming the motifs desired). The following formula describes this last situation:

$$\bigcap_{i=1}^{n_m} \left(\left[\sum_{j=1}^{l_i} PWM_i(b,j) \right]_{\max} \geq \left(PWM_{\max} \right)_i \cdot th \right) \quad (2)$$

Here the variables are the same as described in the previous paragraph. $(\mathrm{PWM}_{\max})_i$ is the maximum position weight score for the i^{th} motif, and *th* (R=[0,1]) is the selection threshold. The reason this is a conjunctive formula is because this implies that there has to be at least one instance of each selected motif in any of the organisms (whose PWM score is greater than the maximum PWM score possible for that motif times the threshold value). In other words each motif has to reach at least the threshold value in order to count as being well-enough formed.

Rate of generation increase

When studying the generation time needed for a given number of motifs we calculated a new measure called the rate of generation increase (RGI). This is given by the following formula:

$$RGI_{ij\ldots k,th} = \frac{G^{nmots}_{ij\ldots k}}{\prod_{i,j,\ldots k} G_i \cdot G_j \cdot \ldots \cdot G_k} \quad (3)$$

where $\mathrm{RGI}_{ij\ldots k,p}$ denotes the generation rate increase for motifs *i, j, …k* at a threshold of *th*. G_i is the generation number for motif *i*, and $G_{ij\ldots k}$ is the generation number for the module constituted of motifs *i, j, …* and *k*, and *nmots* is the number of motifs in the regulatory module being studied. This measure is used to calculate how steeply the formation time increases when more than one motif is studied compared to single motif cases. The RGI value was calculated for 276 motif pairs taken from the SCPD. 24 individual motifs makes (23·24/2)=276 pairs. Since the maximum generation number is 10,000, this means that the minimum RGI value ranges between 0.000001 and 1000000.

List of yeast motifs

The list of the 24 yeast motifs taken from the SCPD database [12] can be seen in Table 1.

Results

Presentation of program

The goal of PromMute is to simulate base pair mutations in a number of separately randomized hypothetical promoter sequences to gauge the speed and dynamics of the formation of information-bearing yeast TFBS's. The number of promoters is set by the user and is the same as the number of hypothetical organisms in the population (since we are studying one single hypothetical gene). Each promoter is 1 Kbp long.

Motif id	Sequence	Motif length	Maximum PWM score
GCR1	CTTCC	5	7.711
GCN4	TGACTC	6	9.482
MCB	ACGCGT	6	9.783
PHO2	TTAAATT	7	9.542
REB1	TTACCCG	7	11.511
SCB	CACGAAA	7	11.323
TBP	TATAAAA	7	10.663
PDR1/PDR3	TCCGCGGA	8	13.546
STE12	ATGAAACC	8	11.469
MATalpha2	CATGTAATT	9	14.137
repressor_of_CAR1	AGCCGCCAA	9	12.798
MCM1	CCCAATTAGG	10	13.914
PHO4	CGCACGTGGT	10	12.144
ABF1	TCACTATACACG	12	14.804
MIG1	CCCCAGATTTTT	12	15.148
RAP1	ACACCCATACAC	12	15.91
ROX1	TCCATTGTTCTC	12	16.124
SWI5	ATATCATGCTGG	12	13.796
UASPHR	TTTTCTTCCTCG	12	14.823
XBP1	GCCTCGAGGCGA	12	15.052
CSRE	TTCGGATGAATGG	13	16.731
GAL4	CGGAGCACACTCCTCCG	17	19.931
RLM1	AGTTCTATAAATAGATTC	18	22.37
SMP1	ATGCTTCTATTTATAGCAAC	20	24.945

Table 1: List of 24 default yeast motifs taken from the SCPD.

Similarly to the ev program, PromMute simulates a mutation/evaluation/ranking/selection/reproduction cycle per generation. A base pair mutation is introduced into each promoter during each generation, the number of which can also be set by the user. Afterwards the best PWM hit is calculated for each selected motif in all promoters. If the best PWM score for all motifs is above the motif selection threshold limit in at least one of the organisms, the simulation stops, since the goal of forming all TFBS's has been achieved. If not, then the promoters are ranked according to their total score value (adding up the best score value for all of the PWM's/motifs), the proportion of organisms with a total best PWM score above the organism selection threshold are selected, and replace those organisms with a lower total best PWM score with another copy of themselves.

Parameters of the PromMute program

On the input screen the following parameters can be set by the user:

- Number of organisms: population size of organisms in the simulation. This is also equal to the number of promoters
- Number of generation cycles
- Motif selection threshold: ratio of position weight matrix score per maximum matrix score for each TFBS
- Organism selection threshold: percent of organisms selected after each generation cycle
- Selected motifs
- Speed: number of generations after which screen is refreshed
- Name of output file

The simulation can be paused and stopped. A screenshot of example input parameters can be seen in Supplementary Figure 1. A short description of the parameters and buttons can be found if the user presses the "Help" tab.

Selection of transcription factor binding sites

By pressing the "Add PWM" button, the user can either manually type in the position weight matrix of a given motif, or can enter data for a seperate motif stored in a seperate file. The motif name and length will then be added to the end of the list of TFBS's on the input screen. Furthermore, the user is given the possibility of adding the PWM of his own TFBS on a seperate tab page, making the program more flexible and realistic.

The number of generation cycles and the number of organisms can be set beforehand. Here the number and kinds of target TFBS's can be selected from a set of 24 experimentally defined TFBS's 5-20 bp long taken from the SCPD, represented by position frequency matrixes. These were converted to position weight matrixes (PWM's) by using the matrix conversion equation given in the convert-matrix tool from RSAT [13]. The program introduces point mutations in each promoter until either the generation limit is met, or all of the selected TFBS's are formed with a PWM score above each individual threshold limit, meaning that the score ratio of the TFBS's PWM score divided by the maximum PWM score are all above the motif selection threshold. During the simulation the individual instances of each of the motifs is colored in various shades of gray. Darker hues mean that the sequence of the specific motif is closer to the target sequence, that is, it's PWM score is almost equal to the maximum PWM score.

Promoter sequence mutation rate

The PromMute program simulates a single base pair mutation per promoter/organism per generation. In the present program a single base pair mutation occurs in each promoter sequence. According to some opinions, a core or proximal promoter sequence (where specific regulation of the promoter takes place) is around 500-2000 bp long [14], which corresponds to a mutation rate of $2 \cdot 10^{-3}$ to $5 \cdot 10^{-4}$, respectively. Taking the spontaneous mutation rate to be around 10^{-8} to 10^{-6}, these rates are clearly too high. Decreasing the mutation rate a technical programming limitation which cannot be avoided. The program could be run with 1000 organisms representing 1000 promoters of the same gene with only one base pair mutation in one of the promoters per

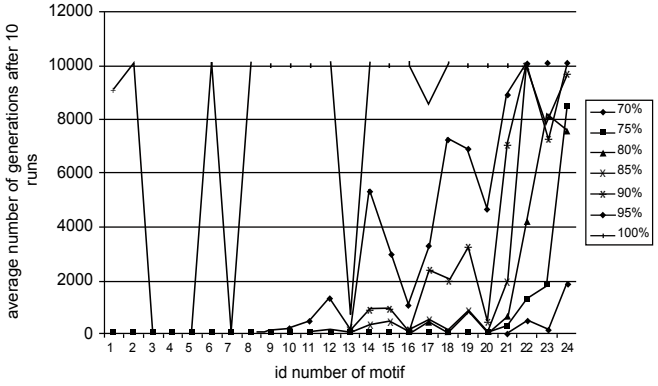

Figure 1: Average number of generations needed for given motif to appear through random mutations. The number of generations needed for each motif to appear through random mutation was simulated 10 times for each motif and then averaged. This was performed for all motif selection threshold values from 0.7-1.0.

generation (which would correspond to a mutation rate of 10^{-6}), but this would slow the simulation down too much. Otherwise, with a larger population more promoters could be available for mutation, but is limited by computer memory.

Selectivity of TFBS's

PromMute applies selection to motifs only after they reach a certain position weight matrix score threshold. The program halts only after all selected motifs have reached the motif selection threshold as supplied by the user. Therefore, similarly to PPE, the program does not preserve partial motif matches.

Program output

Once either the generation limit has been met, or the TFBS network has been formed (with all of the individual TFBS score ratios above the threshold ratio set by the user) the program's output can be viewed on the output tab. Here the highest total sum TFBS score for all promoters is plotted as a function of generation cycle number. The average and lowest total sum TFBS score is also displayed, so the user can get a feel as to how much fluctuation in information content occurs in the formation of TFBS's (see Supplementary Figure 2).

In the output file the input parameters are listed at the head of the file, followed by a list of parameters calculated for each generation of the simulation. This includes the generation cycle number, the best total sum TFBS score, and the position and best PWM score for all selected motifs. This output can be analyzed according to the purposes of the user.

As can be seen from the test output in Supplementary Figure 2, the highest total sum TFBS score fluctuates very much during the simulation. This type of output is similar to the output generated by the ev program of Schneider [9] where selection is not applied. The total lack of information build-up is apparent. This implies that there must be a substantial informational hiatus between random sequences and information-bearing functional sequences such as TFBS's in the promoter which random mutations must be able to hurdle.

Test simulations with prommute

Generation number needed for motif formation: In order to simulate the build-up of genetic information in TFBS's, two sets of simulations were run. In the first setup, the number of generations was calculated which was needed for the formation of every single one of the 24 test TFBS's from yeast. The simulation was performed 10 times and an average generation cycle number was calculated. The maximum generation number was set to 10,000 so the simulation wouldn't go on ad infinitum. The number of organisms was set to 64, and the organism selection threshold to 0.5 (similarly to the parameters used in the ev program).

The average number of generations was plotted for each single TFBS, listed from 1 to 24 for seven different selectivity threshold values (0.7, 0.75, 0.8, 0.85, 0.9, 0.95, and 1.0) (see Table 2). The 24 yeast TFBS's were ranked according to their length, which ranged from 5-20 bp (see Table 1 - the shortest motif was GCR1 with a consensus sequence of CWTCC, and the longest one was SMP1 with a consensus sequence of aWRYTKcTAtWWWTAgMaWY). Therefore the average generation formation time is plotted as a function of motif length as can be seen in Figure 1.

We can see that as a general trend, the generation time for a given

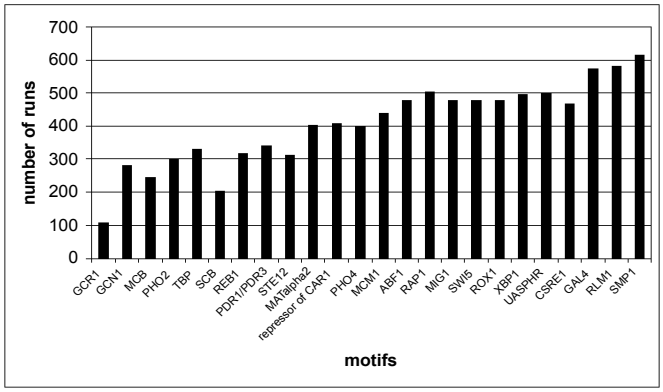

Figure 2: Number of generation runs for every 1000 generations per motif. A test run of 1000 generations was run for each motif at a motif selection threshold of 1.0. The number of runs was counted where the total best PWM score remained the same.

motif increases with the selection threshold and with the length of the motif. Motifs are capable of forming fairly easily under a selection threshold of 0.8. It is quite easy for a motif as long as 13 bp long to form, such as the motif CSRE, taking only 1 generation to form up until a motif thresholf of 90%. However, it takes a longer time for the motifs GAL4, RLM1, and SMP1 to form, which are 17, 18, and 20 bp long, respectively. The corresponding maximum PWM score values for CSRE, GAL4, RLM1, and SMP1 are 16.731, 19.931, 22.37, and 24.945. From this we can infer that TFBS's up to 13 bp long can form quite easily. The situation is different with motifs whose length is between 13 and 17 bp. At a 95% selectivity threshold the program is uncapable of producing motifs longer than or equal to 17 bp, much less at a 100% threshold, where even motifs 13 bp long do not form. Here, only motifs only as long as 7 bp are capable of forming.

Average number of maximum score runs: From Figure 2 we can see that as the motif length and motif selectivity threshold increases, it takes a longer time for a given motif to form. This can be understood, since the longer a motif is, the more likely a back mutation may occur within it. This might happen in spite of a positive mutation occuring in it which would bring it closer sequentially to the target sequence.

A test run of 1000 generations was run for each of the selected 24 yeast motifs at a motif selection threshold of 1.0, and the number of runs was counted where the highest total sum PWM score stayed the same (see Figure 2). That is, a run is broken when a positive or negative mutation occurs in the motif, which either increases or decreases the motif's PWM score (which can be seen in the output tab). The correlation between the length of the motifs and the number of runs is very significant: (r^2=0.922); therefore the longer a motif gets, the easier it is for a mutation to occur in it, giving a larger number of short runs. Furthermore, parallel to this, there is a strong anticorrelation between the motif length and the average run length (r^2=-0.648). This data is presented in Table 3.

Pairs of motifs: We also studied the formation of motif pairs in order to get an impression of how smaller regulatory modules may form through random mutation. To this end we ran the program for all yeast motif pair combinations (276 in total) and counted how many generations were needed for their formation for all motif selection thresholds between 0.7 and 1.0. This information can be seen in the Supplementary Excel file, under the "motif pairs" tab.

	70%	75%	80%	85%	90%	95%	100%
GCR1	1	1	1	1	1	1	9000
GCN4	1	1	1	1	1	1	10000
MCB	1	1	1	1	1	1	1
PHO2	1	1	1	1	1	1	1
TBP	1	1	1	1	1	1	1
SCB	1	1	1	1	1	1	10000
REB1	1	1	1	1	1	1	1
PDR1/PDR3	1	1	3.8	4.1	1	1	10000
STE12	1	1	1	1	1	63.4	10000
MATalpha2	1	1	1	21.3	1	141.5	10000
repressor_of_CAR1	1	1	1	1	1	435.6	10000
PHO4	1	1	1	1	149.8	1257.6	10000
MCM1	1	1	1	1	1	138.6	724
ABF1	1	1	1	294.3	869.9	5360.7	10000
RAP1	1	1	12	383.8	917.5	2920.5	10000
MIG1	1	1.4	85.8	25.4	8.6	1013.5	10000
SWI5	1.3	2.4	438.4	510.4	2371.6	3288.3	8476.4
ROX1	1	1	1	33.6	1959.6	7154.6	10000
XBP1	1	15.4	1	820.4	3206	6850.6	10000
UASPHR	1	1	1	77.8	360.7	4630.8	10000
CSRE1	1	226.3	596.1	1966.2	7006.2	8873.4	10000
GAL4	454.9	1234.7	4227.7	9805.5	10000	10000	10000
RLM1	109.3	1799.1	8174.8	7801.2	7176.4	10000	10000
SMP1	1902.5	8405.6	7491.9	9631.6	10000	10000	10000

Table 2: Average number of generations needed for motif of a given index to form. Rows represent each individual TFBS from the SCPD, whereas columns represent runs done for each motif with a given selectivity threshold (ranging from 70% to 100%).

Motif id	Motif length	Runs/1000 generations	Average run length
GCR1	5	108	9.259259
GCN4	6	279	3.584229
MCB	6	245	4.081633
PHO2	7	297	3.367003
TBP	7	330	3.030303
SCB	7	206	4.854369
REB1	7	317	3.154574
PDR1/PDR3	8	340	2.941176
STE12	8	312	3.205128
MATalpha2	9	403	2.48139
repressor_of_CAR1	9	407	2.457002
PHO4	10	398	2.512563
MCM1	10	439	2.277904
ABF1	12	478	2.09205
RAP1	12	503	1.988072
MIG1	12	477	2.096436
SWI5	12	476	2.10084
ROX1	12	477	2.096436
XBP1	12	495	2.020202
UASPHR	12	498	2.008032
CSRE1	13	467	2.141328
GAL4	17	573	1.745201
RLM1	18	581	1.72117
SMP1	20	614	1.628664

Table 3: Number of runs per 1000 generations and average run length for each of the 24 motifs selected from the SCPD database.

It is interesting to see that when studying motif pairs, the average generation time increases both as a function of motif length and motif selection threshold. At a motif selection threshold of 100%, only very few motif pairs form. The increase in number of generations also increases more and more steeply not only with the motif selection threshold, but with the lengths of the individual motifs in the pair. This can be seen in Figure 3.

For example, if we look at the motif pair GCR1 and GCN4, which are 5 and 6 bp long, respectively, we see that both motifs form almost instantaneously. These two are pairs of short motifs. However, when a small and a medium-sized motif pair is analyzed, such as GCR1 and ABF1 (12 bp long), we find that 109, 309, 766, and 4111 generations are needed for formation at a threshold of 80%, 85%, 90%, and 95%, respectively.

When the short motif GCR1 and the long motif SMP1 (20 bp) are taken into consideration, we can see that the module is incapable of forming at a threshold of 85%. What was also interesting to see in this case was that even at the low motif selection threshold of 0.7, the short 5 bp motif GCR1 changed position on average 551 times, leading to a change of position at least once every 5.1 generations. This is very interesting to note, since it points to the fragility of even very short motifs during the course of molecular evolution. This is a very important factor which one must take into consideration when trying to map the course of molecular evolution since back mutations might occur which tend to break up sequences which have already formed as parts of regulatory modules.

When studying pairs of medium length motifs (motifs ABF1 and MIG1, both 12 bp long), we can see that this motif pair is still capable of forming below a thrshold of 100% (needing 32, 83, 164, 767, 2411, and 5643 generations to form between thresholds of 70%-95%). This means that, although more slowly than pairs of short motifs, formation of a pair of medium length motifs is still fairly rapid. As we can see in Figure 3, the formation profile of this motif pair resembles that of the motifs pairs GCR1-GCN4, and GCR1-ABF1.

However, when we look at a medium length motif paired with a

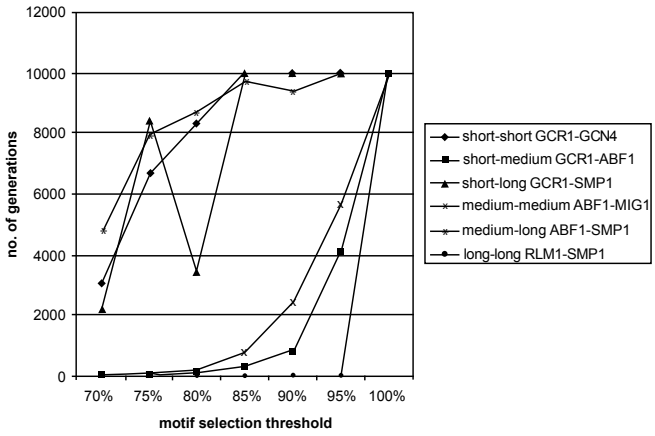

Figure 3: Number of generations needed for pairs of short, medium, and long motifs to form. The number of generations needed for each motif to appear through random mutation was simulated 10 times for a number of select motif pairs and then averaged. This was performed for all motif selection threshold values from 0.7-1.0.

long motif (ABF1 and SMP1, 12 and 20 bp, respectively), we can see that this motif pair reaches the 10000 generation limit at a threshold of 95%, while the long-long motif pair (RLM1 and SMP1, 18 and 20 bp, respectively) reaches this limit at an 85% threshold.

The rate of generation increase was studied for all yeast motif pairs (276 in total). This information is given in the Supplementary Excel file, under the "RGI values" tab. Generation numbers and $\log_{10}$RGI values were depicted as a parameter landscape for motif pairs 1,2 to 1,24 and threshold values from 70% to 100% (the surface of the parameter landscape is similar for all other motif pairs). These can be seen in Supplementary Figures 3 and 4, and the interactive supplementary MatLab figure files pair.fig and rgi.fig. As we can see in the case of the generation time for the motif pairs, there is a steep increase along only a 100% threshold value, as well as for only the longest motifs. However, steep increases in the $\log_{10}$RGI value can be observed even in the middle area of the $\log_{10}$RGI landscape (for example, $\log_{10}RGI_{1,10,0.9}$ = 4.48, where the maximum value is 6.0). (out of the total 276 RGI values, only 24, 8.7% were below 1 with a minimum of $RGI_{3,16,0.85}$ = 0.14). This mirrors the way the generation time needed for multiple motifs in a regulatory module increases as the number of motifs increases.

Discussion

In this article the PromMute promoter mutation simulation program has been presented as a useful tool for researches to study motif turnover and evolution in promoters. The user can select one or more motifs out of 24 yeast motifs to study, or can input their own motif(s). By selecting a motif, the user can study the motif turnover of the specified motif, whereas if many motifs are selected, one can study if a motif regulatory network is capable of being created by random mutation. Furthermore, the sequence, position and highest score can be saved for each generation after each simulation run. By doing this, one can track which mutations lead up to motif formation, as well as studying the dynamic information build-up within a promoter concerning a specific motif or motifs.

Motif appearance and turnover

Acording to the PPE program, Stone and Wray [10] found that 2254 generations were needed for a motif 6bp long to form. This is because the mutation rate applied in their program (10^{-9}) was far lower than that used in PromMute (10^{-3}). According to Lang and Murray [15] when studying two genes in *C. cerevisiae* they calculated the mutation rate to be roughly $5{\cdot}10^{-10}$. However, the population size used in their program was also far larger (10^6 compared to 64, which is the same number used in some test runs in ev). The average waiting time for motif appearance also increased as a function of motif length, 10 times for 7 bp, and 100 times for motifs 8-9 bp long, while it decreased 10-fold for promoters 200 bp long compared to promoters with a length of 2 Kbp (motifs were found on both strands of the DNA).

In comparison, though PromMute started out with a population size of 64, which was roughly 10,000 times smaller than the one in PPE, due to computational restraints, the mutation rate was 10^6 times faster. Since the available space for such a motif is 4 times larger in PPE than PromMute (2 times 2 Kbp, since both strands of the DNA were analyzed), this means that compared to PPE, in PromMute a 6 bp motif appears every 7 generations. As we can see, the yeast motifs GCN4 and MCB, which are both 6 bp long take indeed a very short generation time to form as can be seen in Figure 1 even at high motif selection thresholds. In fact, a motif which is 5 bp long should occur once every 1024 bp, which is approxinately the size of the hypothetical promoter simulated in PromMute, thereby taking in theory also only 1 generation to form.

In order for genes to be expressed properly, they have to undergo fine-tuned regulation, which takes place in the promoter through interaction between a number of TFBS's. For example, in gene regulatory networks, many genes have to integrate signals coming from outside the cell through the interplay of these TFBS's. The question here is, how do these TFBS's form through molecular evolution? As a general tendency we can see that the longer a motif gets, and the larger the number of motifs wich take part in a regulatory module, the longer and more difficult it gets for all of the members of the module to form, let alone taking into consideration the position of the motifs relative to each other. As we can see, a qualitative difference exists in the motif selection threshold-generation time curve when we switch from only one motif to pairs of motifs. However, even pairs of long motifs were able to form after around 8000 generations at a threshold of 85%. This would mean that transcription factors should be able to bind to their respective TFBS's at such a motif threshold value, which is the case in many TFBS's.

Comparison of the program to ev and PPE

The main goal of the PromMute program was to measure the time needed for TFBS's to appear during evolution modelled *in silico* through random base mutations. In this way it is conceptually similar to the program PPE which was specially designed for studying the appearances of regulatory elements restricted to *Drosophila*. Although PPE uses real biological promoters and TFBS, PromMute is able to import new TFBS's supplied by the user besides the 24 yeast motifs used as a default set.

Both PromMute and PPE have a number of new assets compared to ev. Both model real biological data: real TFBS's, of variable lengths, both search for the best fit of a given PWM to a given motif, and allow the motif to form in any part of the promoter. Compared to PPE, PromMute makes it possible for any kind of motif to appear through random mutations, and is not restricted to just one species. Ev suffers from modeling TFBS's from an abstract and purely informatics-

	PromMute	ev	PPE
Length of motifs	5-20 bp	6 bp	5-9 bp
Motifs available from these species	all kinds	hypothetical	7 different species
Mutation model	random	random	stochastic
Mutation, selection, replication cycle per generation	yes	yes	no
Number of occurrence of motifs	1	16	one to a few
Positioning of motif	unrestricted	restricted	unrestricted
Promoter length	1 Kbp	250 bp	200 bp; 2 Kbp
Real promoter sequence	no	no	yes
Real TFBS's used	yes	no	yes
Selection applied to parts of motifs	no	yes	yes
TFBS matching and scoring	PWM	two's complement	PWM

Table 4: List of similarities and differences between PromMute, ev, and PPE.

centered viewpoint. Here motifs are coded in binary, and mutations occur according to how the binary form of a motif is changed.

One of the fundamental differences between PromMute and ev however is the way selection is applied to forming motifs. Both programs simulate a set number of hypothetical organisms which undergo a cycle of mutation, evaluation, selection, and reproduction, however, the real question is how natural selection acts on the motifs under study. The ev program retains fractions of motifs which may be half-formed. In comparison PromMute accepts a motif only if it has reached a certain motif threshold value. That is, it accepts a fit between a TFBS and its TF if it is at least "almost good", sort of the way a key fits into a lock, albeit with a minimum amount of wobbling allowed. Until the motif is formed, the motif undergoes "sequential drift" allowing all kinds of base pair mutations to undergo within the motif. As seen in Figure 1, a good candidate threshold would be 80%, since at 85% the longest motifs are still not capable of forming. Indeed, according to a study by Collado-Vides et al. [16], different sites for the same binding motif differ from each other by 20-30% in *E. coli*. Table 4 presents the similarities and differences of the PromMute, ev, and PPE discussed above.

Outlook

More features can be added to the program, thereby making it more realistic. For example, the program may start with promoter sequences containing real TFBS's. In this scenario the goal of the program could be changed to see whether a new kind of TFBS could appear through random mutations to augment an already existing TFBS network. The length of hypothetical promoter sequences could also be set by the user, since promoters can vary in length according to their function. Mutational hot spots could be taken into account. Furthermore, besides basic base pair mutations, larger mutations such as insertions or deletions can also be modelled.

Overall, the PromMute program is a multi-faceted simulation program which can aid research by visualizing sequential changes in promoters step by step and mutation by mutation. The reason for using it to model molecular evolution is it's easy usage, speed, usage of real motifs, and it's capability of producing data for further analysis.

Acknowledgements and Funding

The author would like to thank György Süveges from the University of Szeged, Faculty of Natural Sciences and Informatics (SZTE-TTIK) for his valuable technical help in writing the program and Tamás Turányi for critically reading the paper. The European Union and the European Social Fund have provided financial support to the project under the grant agreement no. TÁMOP 4.2.1./B-09/1/KMR-2010-0003.

References

1. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, et al. (2003) The Evolution of Transcriptional Regulation in Eukaryotes. Mol Biol Evol 20: 1377-1419.
2. Ohno S (1970) Evolution by Gene Duplication. Springer-Verlag, Heidelberg.
3. Kim JT, Martinetz T, Polani D (2003) Bioinformatic principles underlying the information content of transcription factor binding sites. J Theor Biol 220: 529-544.
4. Erill I, O'Neill MC (2009) A reexamination of information theory-based methods for DNA-binding site identification. BMC Bioinformatics 10: 57.
5. Moses AM (2009) Statistical tests for natural selection on regulatory regions based on the strength of transcription factor binding sites. BMC Evol Biol 9: 286.
6. Shultzaberger RK, Roberts LR, Lyakhov IG, Sidorov IA, Stephen AG, et al. (2007) Correlation between binding rate constants and individual information of *E. coli* Fis binding sites. Nucleic Acids Res 35: 5275-5283.
7. Schneider TD, Stormo GD, Gold L, Ehrenfeucht A (1986) Information content of binding sites on nucleotide sequences. J Mol Biol 188: 415-431.
8. Schneider TD (1997) Information Content of Individual Genetic Sequences. J Theor Biol 189: 427-441.
9. Schneider TD (2000) Evolution of biological information. Nucleic Acids Res 28: 2794-2799.
10. Stone JR, Wray GA (2001) Rapid evolution of cis-regulatory sequences via local point mutations. Mol Biol Evol 18: 1764-1770.
11. Zhu J, Zhang MQ (1999) SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. Bioinformatics 15: 607-611.
12. SCPD The Promoter database of Saccharomyces cerevisiae.
13. Thomas-Chollier M, Sand O, Turatsinze JV, Janky R, Defrance M, et al. (2008) RSAT: regulatory sequence analysis tools. Nucleic Acids Res 36: W119-W127.
14. Lichtenberg J, Yilmaz A, Welch JD, Kurz K, Liang X, et al. (2009) The word landscape of the non-coding segments of the *Arabidopsis thaliana* genome. BMC Genomics 10: 463.
15. Lang GI, Murray AW (2008) Estimating the per-base-pair mutation rate in the yeast *Saccharomyces cerevisiae*. Genetics 178: 67-82.
16. Collado-Vides J, Magasanik B, Gralla JD (1991) Control site location and transcriptional regulation in *Escherichia coli*. Microbial Rev 55: 371-394.