Locating potential enhancer elements by comparative genomics using the EEL software

Kimmo Palin[1], Jussi Taipale[2] & Esko Ukkonen[3]

[1]Department of Computer Science, P.O. Box 68 (Gustaf Hällströmin katu 2b) FIN-00014, University of Helsinki, Finland. [2]Molecular/Cancer Biology Program, Institute of Biomedicine, University of Helsinki and Department of Molecular Medicine, National Public Health Institute (KTL), Biomedicum, P.O. Box 63 (Haartmaninkatu 8), FIN-00014, University of Helsinki, Finland. [3]Helsinki Institute for Information Technology, P.O. Box 68, FIN-00014, University of Helsinki, Finland. Correspondence should be addressed to K.P. (Kimmo.Palin@helsinki.fi).

Published online 27 June 2006; doi:10.1038/nprot2006.56

This protocol describes the use of Enhancer Element Locator (EEL), a computer program that was designed to locate distal enhancer elements in long mammalian sequences. EEL will predict the location and structure of conserved enhancers after being provided with two orthologous DNA sequences and binding specificity matrices for the transcription factors (TFs) that are expected to contribute to the function of the enhancers to be identified. The freely available EEL software can analyze two 1-Mb sequences with 100 TF motifs in about 15 min on a modern Windows, Linux or Mac computer. The output provides several hypotheses about enhancer location and structure for further evaluation by an expert on enhancer function.

INTRODUCTION

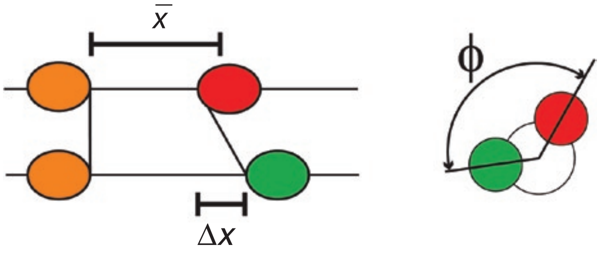
Locating distal enhancer elements in mammalian genomes is difficult because of the vast amount of DNA in which the elements can be situated and because of the variable conservation of the elements themselves. The computer program EEL[1] locates putative enhancer elements within long stretches of genomic DNA. As input, EEL requires two orthologous DNA sequences and the position-specific binding profile matrices for a suitable set of TF binding motifs.

The EEL algorithm locates the highest-energy enhancer elements according to a simplified biochemical and physical model of TF binding (**Fig. 1**). Briefly, the TF complex that binds to the enhancer element gains energy by binding the TFs to their respective binding sites. Additional energy is gained by secondary interactions (i.e., by association of the TFs with one another directly or through other proteins during formation of the complex that activates transcription at the promoter). In the EEL model, the secondary interaction energy is assumed to be directly proportional to the DNA-binding energy of the TFs. Increased distances between TF sites decrease the likelihood and expected energy contribution of the secondary interactions. EEL takes this into account by making the energy gained from the secondary interactions inversely dependent on the distance between the adjacent binding sites. Secondary interactions are also less likely to be conserved if the TFs that bind to two sequences are at different relative distances and different angles with respect to one another. In such cases, EEL estimates the energy required to generate similar spatial positions of the TFs in both species by compressing and twisting the DNA helix between them.

The energy equation is composed of four terms that are weighted by parameters Lambda, Mu, Nu and Xi (which default to 2.0, 0.5, 200.0 and 200.0, respectively). These parameters control the relative contributions to the EEL score of (i) the binding affinity of the TFs to their respective binding sites (Lambda), (ii) the distance between adjacent binding sites (Mu), (iii) the difference in the distance of the TF sites between the two species (Nu) and (iv) the difference in the angle of the TFs between the two species (Xi). The minimum value of each parameter is 0, which leads to the elimination of the influence of the respective term.

Fundamentally, the EEL software finds conserved clusters of binding sites on two DNA sequences. Thus, there is no reason for the DNA-binding proteins to be limited to TFs; EEL can be used to locate any kind of element within DNA that contains short conserved motifs that have representation as position-specific scoring matrices separated by more freely evolving sequences[2].

Enhancer prediction tools can be characterized with respect to their required input and the underlying enhancer model. The



$$\text{Score} = \lambda\Delta G - \mu\bar{X} - \frac{\nu\Delta X^2 - \xi\phi^2}{2\bar{X}}$$

Figure 1 | The scoring function of EEL. Binding sites with affinity ΔG are conserved in human and mouse sequences. The average distance to adjacent binding sites in both species is $\bar{x}$, and the difference in this distance in the two species is Δx. Along the distance Δx, the DNA helix turns ϕ radians, changing the location of the binding site in one of the species. The local alignments that result in the enhancer predictions are extended and scored according to the equation shown. Briefly, the score of an enhancer element is proportional to the energy of binding of the TF complex to the enhancer. The parameters *Lambda* (λ), *Xi* (ξ), *Nu* (ν) and *Mu* (μ) can be used to adjust the contribution of the different terms of the scoring function to the score. Increasing λ results in a higher contribution of the DNA binding affinity of the TFs to the overall score. Increasing μ results in a higher penalty for the distance between adjacent TF sites, which makes a close clustering of the TF sites more important. Increasing ν and ξ increases the importance of the conservation of the distances and angles, respectively, between adjacent TFs. All of the parameters must be positive. (For additional details, see ref. 1 and the online supplemental material thereof.)

biologically, though not computationally, simplest methods use only phylogenetically or functionally related DNA sequences to locate either individual binding sites[3], ultra-conserved noncoding regions[4] or regions with abnormal sequence composition[5–7]. Common to these methods is that they cannot identify the TFs that regulate the located sites.

More-sophisticated methods also require the binding site motifs of some TFs as inputs and model the enhancers combinatorially[8]. This class covers a wide range of tools from simple frequency-counting methods[9] to methods that model the structure[10] and possibly the conservation[11–14] of the enhancers. EEL is also of this type. Typically, these methods are fast, and the results are easy to interpret in terms of the bound TFs.

A major class of enhancer prediction tools is the probabilistic methods[8]. These tools use Bayesian networks[15] or hidden Markov models[16–19] and benefit from the well-developed theory and algorithms of Bayesian statistics. The tools normally indicate the regions with the maximal likelihood of being an enhancer with respect to their enhancer model. Probabilistic modeling leads to computationally heavy algorithms, which make these tools quite slow in practice.

For example, 'Stubb'[17], which may be the most sophisticated probabilistic enhancer-locating algorithm, can take into account two orthologous sequences, and the underlying probabilistic model can handle several phylogenetically related species. Stubb can also model the dependencies between TFs. We found that Stubb requires 45 min to analyze a single sequence (a gene with 100 kb flanking sequences with exons and tandem repeats masked and 107 binding motifs; multiple-sequence analysis fails with Stubb 2.1), whereas EEL analyzed two such orthologous sequences in 2 min. The quality of the Stubb results is sensitive to the number and similarity of the binding-site motifs[17], as "weak motif occurrences also contribute to the score" and as overfitting may become a problem with motif sets that are larger than about 20 (ref. 17). We conclude that EEL and Stubb are designed to achieve different goals. EEL is better suited to the exploratory analysis of long sequences with a large collection of binding motifs, whereas Stubb is better for analyzing short sequences with relatively well-characterized regulatory functions.

The main advantages of EEL are as follows:

1. It has a general high-throughput nature, which permits the simultaneous analysis of large numbers of TFs and the generation of databases of predicted enhancers and transcription factor binding sites (TFBSs). Thus, specific hypotheses can be tested after the general analysis using database queries, which are much less computationally intensive than the alignment process itself.
2. It is applicable to very long sequences.
3. Its enhancer model is based on physical interactions that occur within the enhancer.

The main limitations of EEL are as follows:

1. It finds only enhancers that are composed of TFBSs of TFs whose binding-specificity matrices are provided to the program.
2. It is based on an analysis of the order of the TF sites, and thus it cannot find enhancer elements where TFBSs are conserved but their order is not. This may make EEL less suitable for the analysis of distantly related species.
3. A good understanding of the structure of enhancers and average scores that are obtained using the TF sites and species tested is required to interpret the output of EEL.

EEL software requires prior knowledge about the binding affinity matrices of TFs. As more information about binding affinities becomes available, however, this limitation is becoming less of an issue. Indeed, we have found that the software works well with our current subset of binding affinities for approximately 100 TFs. As the binding sites are clustered, not all of the binding affinities need to be known; the enhancer will be located by EEL if a high enough proportion of the conserved sites of the cluster are for those TFs whose affinity matrices are known. Thus, although EEL is not designed to find new TF motifs[3,20,21], such motifs could potentially be identified by analyzing the DNA sequences of the predicted enhancer elements.

It should also be noted that because EEL is a comparative tool, it finds only potential enhancers that are conserved in both of the species analyzed: many of the nonconserved enhancers are needed to regulate nonconserved patterns of gene expression. Although this clearly limits the utility of EEL, it also increases its specificity[22]. Another limitation of EEL is that it is not capable of finding enhancers in highly conserved regions of DNA. For example, comparison of the human and mouse HOX gene clusters using the default parameter values of EEL reports the whole sequence as one big enhancer. This is due to the high conservation of the vertebrate HOX gene clusters. Also, tandem repeats and coding regions obscure the results by providing high-scoring clusters of binding sites that clearly differ from the typical genomic landscape but are not likely to have regulatory functions.

Although EEL can locate putative enhancer elements without a prior knowledge of which TFs regulate the target gene, a basic understanding of the real enhancer and the TFBSs that the sequence is expected to contain should be applied when selecting predicted enhancer elements for further wet-lab study.

The use of EEL is straightforward. First, an orthologous pair of DNA sequences and a collection of TF motifs are selected for analysis. Then the computer searches for putative TFBSs in the provided sequences and aligns the conserved sites with respect to the energy-based scoring function (**Fig. 1**). The final results are given in decreasing order with respect to the score, and they can be observed either visually or saved to a file for further processing.

The orthologous DNA sequences should be derived from species that are evolutionarily close enough so that the enhancer elements are conserved, but are distant enough so that the intervening sequences have diverged. For the human genome, we have found that chimpanzee sequence is generally too closely related, whereas comparisons to mouse, rat or dog seem to yield good results. Nonmammalian vertebrate sequences, in turn, are so divergent from human sequences that the sensitivity of EEL decreases[1]. This is partly because of divergent gene regulation and partly because EEL requires the order of the TFBSs to be conserved, a requirement that is made for computational efficiency. In general, using more-distant species should result in increased specificity but decreased sensitivity. Also, when a more conserved biological process is studied, species that are more divergent can be used in the analysis. For example, in identifying enhancers regulated by signaling pathways

Figure 2 | A file containing the position-specific binding profile for the TF Hunchback obtained from the JASPAR database. Each column describes the nucleotide distribution for one position in the binding site. The rows from top to bottom stand for nucleotides A, C, G and T, respectively. For example, half (8 of 16) of the observed binding sites for Hunchback contain a C at their second position.

1	6	9	4	13	16	16	14	15	9
5	8	3	3	1	0	0	0	1	2
8	2	4	1	0	0	0	2	0	2
2	0	0	8	2	0	0	0	0	3

that regulate early embryonic development, useful results can be obtained by comparing human sequence to puffer fish sequence (see Supplementary Table S6 in ref. 1).

The EEL software also provides command line and console interfaces. These features allow a computer expert to write scripts for running EEL on large sets of orthologous genes, possibly in parallel. The list of command line options with brief descriptions is displayed by starting EEL with command eel -help. The console interface is used when starting EEL with option -nogui, and the command help is provided with the command help.

MATERIALS

EQUIPMENT

- Modern personal computer (see EQUIPMENT SETUP)
- EEL software (see EQUIPMENT SETUP)

EQUIPMENT SETUP

- **Modern personal computer** The operating system should be Windows (Microsoft) or Linux (e.g., RedHat or SUSE). Other UNIX-like operating systems will probably also work fine. The software for MacOS X (Apple) is provided but is not supported, and the graphical user interface is significantly different from that used by the other systems.
- **EEL software** The EEL software is available at http://www.cs.helsinki.fi/u/kpalin/EEL/ under GNU GPL license (http://www.gnu.org/copyleft/gpl.html). Select, download and install the appropriate version of the software by following the on-screen instructions. The binding site motifs used in the analysis are represented as text files with four rows of integer numbers and columns separated by space or tab characters. An example file representing a typical position-specific binding profile matrix is shown in **Figure 2**. The numbers are the counts or frequencies of each particular nucleotide in the column. The first through fourth rows stand for the nucleotides in alphabetical order, A, C, G and T, respectively. Such DNA-binding motif matrices are available from databases such as JASPAR[23] (http://mordor.cgb.ki.se/cgi-bin/jaspar2005/jaspar_db.pl), or they can be obtained by laboratory procedures[24] or by computational studies[3,20,21]. Finally, EEL needs two orthologous DNA sequences from two species; these are searched for enhancer elements and must be provided as files in the FASTA (http://www.ncbi.nlm.nih.gov/blast/fasta.shtml) format. Users can use their own sequences or fetch them from a sequence database or genome browser such as ENSEMBL[25] (http://www.ensembl.org/) or UCSC genome browser[26] (http://genome.ucsc.edu/). The sequences can be up to a few million base-pairs long depending on the available memory of the computer. The regions of the input sequence that the user wants to exclude from the analysis can be masked before the EEL analysis by replacing the original sequence symbol with 'N' according to sequence annotation or tools such as Tandem Repeats Finder[27] or RepeatMasker (http://www.repeatmasker.org).

PROCEDURE

1| To start EEL, click Start→All Programs→Enhancer Element Locator→EEL (on Windows Server 2003) or by giving command eel in the terminal window. A gray window with the title 'Enhancer Element Locator' is displayed (**Fig. 3**).

2| To add sequences, click the 'Add Sequences' button. A window such as the one shown in **Figure 4** opens. Select the FASTA-formatted files that contain the orthologous sequences that are to be searched for putative enhancers. When the sequences have been added, click 'OK'. The 'Add Sequences' screen will close, leaving only the main window showing. The names of the DNA sequences that are contained in the files will be shown in the left-hand list of the main window (**Fig. 4**). A sequence can be removed from the analysis by double clicking its name on the list in the main window. In the end, the list should contain the names of the two sequences to be searched for enhancers.

▲ **CRITICAL STEP** Check the sequences before loading using a sequence analysis tool or text editor to determine whether they contain ambiguous nucleotides because of missing sequence. Sequences that contain the symbol N (representing A, C, G or T) can lead to false-negative results because EEL does not allow N to overlap with a TF site.

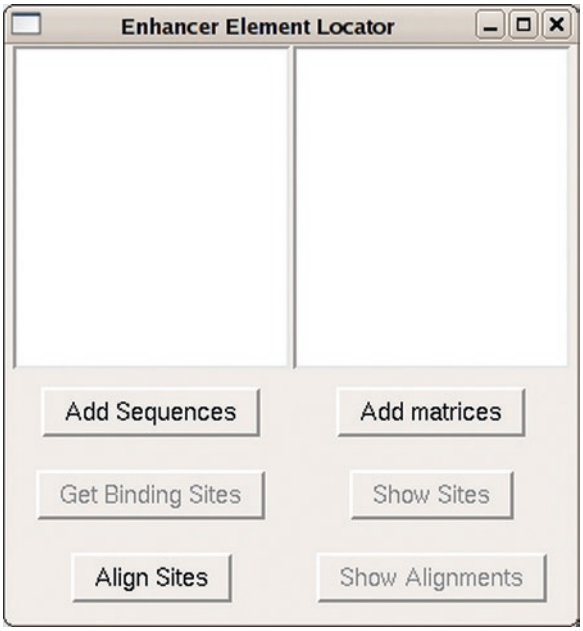
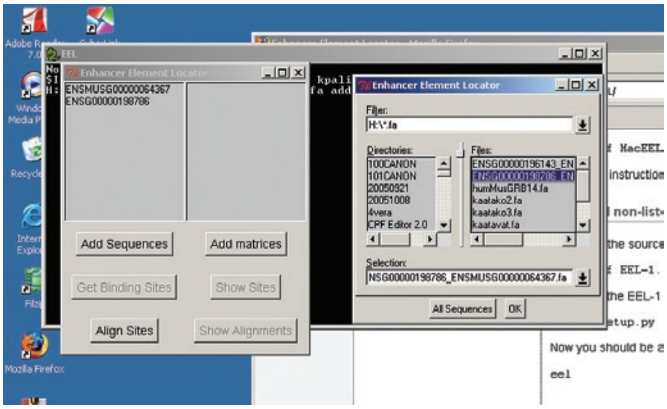


Figure 3 | The main window of EEL when the program is started. On the top left, there is a space where the sequences will be listed, and on the right there is a space for the binding motif matrices to be used in the analysis. Below the list areas, there are command buttons, some of which are initially disabled.

3| To add binding motif matrices, click the 'Add matrices' button. The window in **Figure 5** appears, which enables the selection of files containing DNA-binding motifs to be used in the analysis. When all binding matrices are selected, click 'OK'. The 'Add matrices' window will close, leaving only the main window, which will now contain a list of the sequences (from Step 2) and a list of matrices. To decrease the statistical error, the matrices should optimally be derived from direct affinity measurements[24] or from alignment of a relatively large number of binding site sequences (e.g., the JASPAR[23] database contains such high-quality matrices). Optimally, one should include in the analysis only the motifs that constitute the enhancer(s) of interest. As this set is typically unknown, it is better to include a set of motifs that are potentially involved. If no prior knowledge exists, or if a general analysis is desired, all available motifs should be included. The minimum number of motifs is one. The names of the binding matrix files, along with the information content of the matrix, will be shown in the right-hand list of the main window (**Fig. 5**), and the matrices can be removed by double clicking.

4| To search for the TFBSs, click the 'Get Binding Sites' button. The 'Get TFBS...' window shown in **Figure 6** opens. Select the background distribution assumption (i.e., for the non–binding site coding sequence) and the cutoff for motif matching. Finally, click 'Get TFBS' and wait while the computer searches for the TFBSs. This might take a few minutes depending on the length and number of the sequences and the number of binding matrices considered. The "Get TFBS..." window will close once the search is complete, leaving only the main window, with the 'Show Sites' and 'Align sites' buttons now enabled. The available options for the background assumption are either an independent, identically distributed background ('Uniform background') with arbitrary base composition or a prelearned Markov background, which depends on a short DNA context before the modeled nucleotide. By using a Markov model of local nucleotide dependencies, the binding matrix scores can be corrected for the fact that some short sequences (e.g., AT-repeats and CpGs) are found too frequently or rarely in the genome. Users can make their own Markov background distributions within the console or command-line interface with the commands setMarkovBG and saveMarkovBackground. A fourth-order Markov background learned from human chromosome I is provided in the EEL folder in the file human.ChrI.04.bg. The putative TFBSs (i.e., the binding matrix matches) are scored as log-odds between the motif and background likelihoods. The absolute cutoff is the lower bound on these log-odds scores and is fixed for all binding matrices. The relative cutoff is the fraction of the maximum score attainable for a binding matrix that is still accepted as a valid TFBS. Relative cutoffs are only applicable to the uniform background assumption, where the maximum score of a binding matrix is unambiguously defined.



Enhancer Element Locator

ENSMUSG00000064367
ENSG00000198786

Enhancer Element Locator

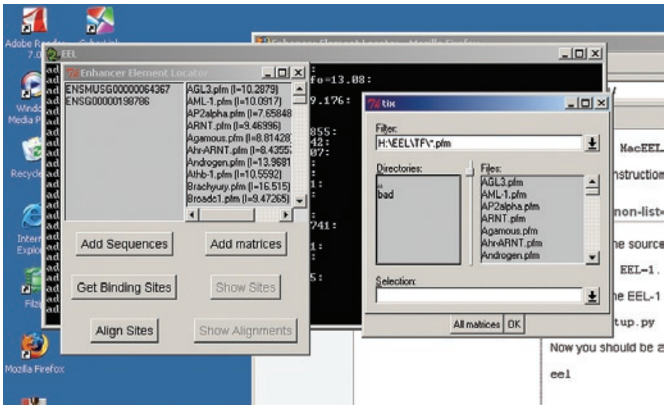
Filter: H:\*.fa

Directories: 100CANON, 101CANON, 20050921, 20051008, 4vera, CPF Editor 2.0

Files: ENSG00000196143_EN, ENSG00000198786_EN, humMusGRB14.fa, kaatako2.fa, kaatako3.fa, kaatavat.fa

Selection: NSG00000198786_ENSMUSG00000064367.fa

Figure 4 | Display shown during addition of new sequences to the analysis. The names of the added sequences show up on the left-hand list of the main window. By clicking 'All Sequences' button, the sequences in all the files shown in the 'Files:' list are added to the analysis.



Enhancer Element Locator

ENSMUSG00000064367
ENSG00000198786

AGL3.pfm (I=10.2879)
AML-1.pfm (I=10.0917)
AP2alpha.pfm (I=7.65848
ARNT.pfm (I=9.46996)
Agamous.pfm (I=8.81428
Ahr-ARNT.pfm (I=8.4355
Androgen.pfm (I=13.9681
Athb-1.pfm (I=10.5592)
Brachyury.pfm (I=16.515)
Broadc1.pfm (I=9.47265)

tix

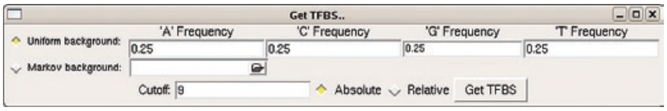
Filter: H:\EEL\TF\*.pfm

Directories: .., bad

Files: AGL3.pfm, AML-1.pfm, AP2alpha.pfm, ARNT.pfm, Agamous.pfm, Ahr-ARNT.pfm, Androgen.pfm

Selection:

Figure 5 | Display shown during addition of new binding motif matrices to the analysis. The names of the binding motif files and the information content of the motifs are shown on the right-hand list of the main window. By clicking the 'All matrices' button, all motifs shown in the 'Files:' list can be added to the analysis.



Get TFBS..

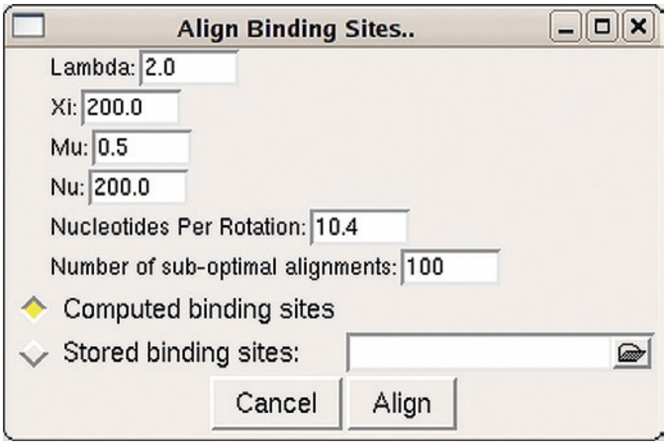
	'A' Frequency	'C' Frequency	'G' Frequency	'T' Frequency
● Uniform background:	0.25	0.25	0.25	0.25
○ Markov background:				

Cutoff: 9 ● Absolute ○ Relative

Figure 6 | The window for parameters of the TFBS search. The user can provide either an independent, identically distributed background distribution or a Markov background distribution for the search. The cutoff for the binding motif match score can be provided either as an absolute score or, optionally, all binding motif matches can be considered relative to their respective maximum score. The latter option is available only if the uniform background distribution is used.

5| At this point, the identified binding sites can be saved for subsequent analysis by EEL or other software. Click 'Show Sites' on the main window to open the 'Putative TF binding sites' window. Select 'Save' and enter the name of the file where the list of sites should be saved. The list will be formatted in the General Feature Format (GFF; http://www.sanger.ac.uk/Software/formats/GFF/), which can be used later by EEL or as an input to many other software tools.

6| Align the identified binding sites with EEL. On the main window, click the 'Align Sites' button. The 'Align Binding Sites...' window will open (**Fig. 7**), allowing the parameters (**Fig. 1**) for locating the conserved enhancer elements to be set. The default values for parameters Lambda, Xi, Nu, Mu and Nucleotides Per Rotation are 2.0, 200.0, 200.0, 0.5 and 10.4. Finally, click 'Align'. If binding sites from more than two sequences were provided, a new window will pop up asking which two sequences should be used in the alignment. The alignment should not take more than a few minutes. The main window with all command buttons enabled will reappear.



Align Binding Sites..

Lambda: 2.0

Xi: 200.0

Mu: 0.5

Nu: 200.0

Nucleotides Per Rotation: 10.4

Number of sub-optimal alignments: 100

◆ Computed binding sites

◇ Stored binding sites:

Figure 7 | A window for parameters of the local alignment procedure for enhancer prediction. The parameters *Lambda*, *Xi*, *Mu*, *Nu* and *Nucleotides Per Rotation* adjust the scoring of the enhancers (**Fig. 1**). The number of suboptimal alignments is the number of putative enhancer elements that will be reported as output. The user can choose to align the binding sites that are in memory or that were previously stored in a GFF-formatted file in Step 5.

7| After carrying out an EEL alignment, click on the 'Show Alignments' button on the main window to view the results. A window like that shown in **Figure 8** is displayed. The putative enhancer elements are displayed from most likely to least likely. The output can be saved to a file by clicking the 'Save human readable' button. By clicking 'Save computer readable', the user can save the results in GFF format, which is better suited for further computational analyses. The window shown in **Figure 8** stays open after the output is saved in either format. Click OK to close the window.

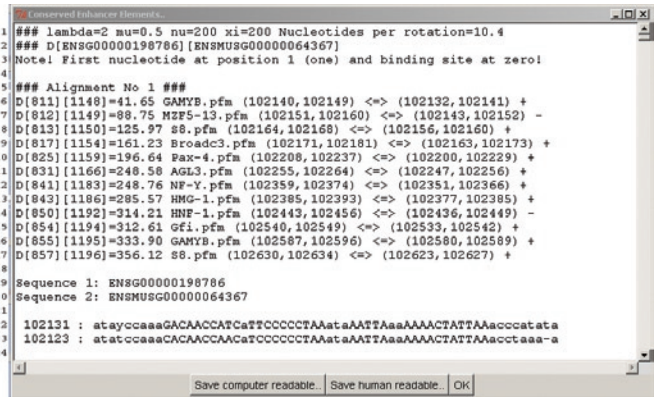
? TROUBLESHOOTING

The default parameter values for EEL scoring should be applicable[1] for comparing most human sequences to sequences from mouse and other mammals. The default parameter values are not appropriate for unusually well-conserved regions or for comparing nonmammalian species. In these cases, the user must resort to a trial-and-error procedure to find parameter combinations that produce reasonable results for that particular case. The parameters can be reoptimized, for example, by changing them systematically and comparing the scores for a number of known enhancers to average scores.

In general, increasing Xi and Nu will make the predicted enhancers better conserved, and increasing Mu will make the TFBSs in the enhancers closer together. Increasing any of the parameters Xi, Nu or Mu is likely to make the predicted enhancers shorter. Lambda is relative to all the other parameters and can be left fixed at all times.

ANTICIPATED RESULTS

After running EEL, the user should have good hypotheses about where and what kind of enhancer elements, if any, are located in the sequences. It should be stressed that the output produced by EEL indicates hypothetical TFBSs and enhancer elements, and these predictions should be experimentally validated. When predicted elements are selected for experimental validation, the EEL score of the element, its length, the binding sites that it contains and



```
 1 ### lambda=2 mu=0.5 nu=200 xi=200 Nucleotides per rotation=10.4
 2 ### D[ENSG00000198786][ENSMUSG00000064367]
 3 Note! First nucleotide at position 1 (one) and binding site at zero!
 4
 5 ### Alignment No 1 ###
 6 D[811][1148]=41.65 GAMYB.pfm (102140,102149) <=> (102132,102141) +
 7 D[812][1149]=88.75 MZF5-13.pfm (102151,102160) <=> (102143,102152) -
 8 D[813][1150]=125.97 S8.pfm (102164,102168) <=> (102156,102160) +
 9 D[817][1154]=161.23 Broadc3.pfm (102171,102181) <=> (102163,102173) +
10 D[825][1159]=196.64 Pax-4.pfm (102208,102237) <=> (102200,102229) +
11 D[831][1166]=248.58 AGL3.pfm (102255,102264) <=> (102247,102256) +
12 D[841][1183]=248.76 NF-Y.pfm (102359,102374) <=> (102351,102366) +
13 D[843][1186]=285.57 HMG-1.pfm (102385,102393) <=> (102377,102385) +
14 D[850][1192]=314.21 HNF-1.pfm (102443,102456) <=> (102436,102449) -
15 D[854][1194]=312.61 Gfi.pfm (102540,102549) <=> (102533,102542) +
16 D[855][1195]=333.90 GAMYB.pfm (102587,102596) <=> (102580,102589) +
17 D[857][1196]=356.12 S8.pfm (102630,102634) <=> (102623,102627) +
18
19 Sequence 1: ENSG00000198786
20 Sequence 2: ENSMUSG00000064367
21
22  102131 : atayccaaaGACAACCATCaTTCCCCCTAAataAATTAaaAAAACTATTAAacccatata
23  102123 : atatccaaaCACAACCAACaTCCCCCCTAAataAATTAaaAAAACTATTAAacctaaa-a
24
```

Figure 8 | The window showing the predicted enhancer elements. Line 1 shows the parameter settings used in the alignment, and line 2 gives the names of the aligned sequences. Lines 6–17 show the actual alignment: indexes of the aligned sites in the two sequences, the score of the alignment up to this pair of sites, name of the TF binding to these sites, begin and end positions of the binding sites in the two sequences and the strand of DNA where the sites are found (plus for Watson's strand, minus for Crick's). The total score of the enhancer element is on line 17. Beginning at line 22 is a display of the alignment of the underlying DNA, if the sequence is available from Step 2. Although the DNA between the binding sites is aligned, only the binding sites (shown in uppercase characters) are used in finding the conserved enhancer elements.

the potential presence of repetitive and coding sequences should all be considered.

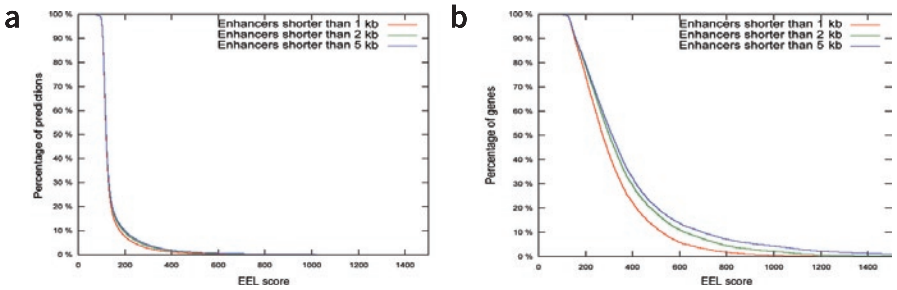


Figure 9 | The score distributions of genome-wide EEL alignments for human and mouse. (**a**) The percentage (*y* axis) of predicted enhancers of maximum length 1, 2 or 5 kb with the indicated or higher score (*x* axis). For each pair of genes, 50 enhancers were predicted. For example, about 8% of enhancers shorter than 1 kb score better than 200. (**b**) Percentage of all analyzed genes (*y* axis) that have a predicted enhancer of maximum length 1, 2 or 5 kb with the indicated or higher EEL score (*x* axis). For example, about 10% of human genes have an enhancer prediction shorter than 1 kb with an EEL score of 500 or higher. EEL alignments were carried out using the default parameters and 107 TFBSs; all orthologous human and mouse gene sequences, starting from 100 kb upstream of the transcription start site and ending 100 kb downstream from the end of the last exon, were used to generate the scores (for details, see ref. 1).

The enhancer model used in EEL is based on a biochemical and physical model of TF binding to DNA, and because of this, EEL, unlike many other alignment tools, does not provide *P* values for its predictions. Evaluating the quality of the predictions is best carried out by estimating the plausibility of the overall structure of each of the enhancers identified, and by comparing the scores to those obtained from analyses of known enhancer elements using the same parameters and species. The quality of the predicted enhancers can also be analyzed by comparing the obtained scores to the distribution of all scores from a genome-wide analysis. The distribution of EEL scores of all enhancers with lengths less than 1, 2 and 5 kb from a genome-wide human-mouse comparison using the default parameters (from a similar analysis as described in ref. 1) is shown (**Fig. 9a**). The percentage of genes that contain at least one predicted enhancer as a function of the EEL score used as a cutoff is also shown (**Fig. 9b**). For default parameter settings, when comparing human and mouse sequences, scores above 500 can be considered significant. The scores are generally lower when more distantly related species are compared, and thus a lower cutoff for significance can be used (e.g., 150 for human to puffer fish and 250 for human to chick). Additional properties that should be considered are the length of the enhancer, which generally should not be more than 2 kb, and the TF sites that comprise the enhancer. It is a good sign if the enhancer contains a site for a factor that is known to regulate the gene of interest.

Not all predicted enhancers will warrant further investigation. For example, tandem repeats can generate high-scoring alignments that are comprised of sites for one or a few factors. Regions of the genome, such as coding sequences, which are highly conserved because of factors that are not related to gene regulation, can also generate anomalous high-scoring alignments. Thus, alignments in known coding or repetitive regions are more likely to be false positives than are alignments from nonrepetitive and/or noncoding sequences.

ACKNOWLEDGMENTS This work was supported by the 'Translational Genome-Scale Biology' and the 'From Data to Knowledge' Centers of Excellence of the Academy of Finland and by the 'BioSapiens' and the 'Regulatory Genomics' projects of the European Union.

COMPETING INTERESTS STATEMENT The authors declare that they have no competing financial interests.

Published online at http://www.natureprotocols.com
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Hallikas, O. *et al.* Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* **124**, 47–59 (2006).
2. Cameron, R.A. *et al.* An evolutionary constraint: Strongly disfavored class of change in DNA sequence during divergence of cis-regulatory modules. *Proc. Natl. Acad. Sci. USA* **102**, 11769–11774 (2005).
3. Tompa, M. *et al.* Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.* **23**, 137–144 (2005).
4. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321–1325 (2004).
5. Nazina, A.G. & Papatsenko, D.A. Statistical extraction of *Drosophila* cis-regulatory modules using exhaustive assessment of local word frequency. *BMC Bioinformatics* **4**, 65 (2003).
6. Grad, Y.H., Roth, F.P., Halfon, M.S. & Church, G.M. Prediction of similarly-acting cis-regulatory modules by subsequence profiling and comparative genomics in *D. melanogaster* and *D. pseudoobscura*. *Bioinformatics* **20**, 2738–2750 (2004).
7. Segal, E. & Sharan, R. A discriminative model for identifying spatial cis-regulatory modules. *J. Comput. Biol.* **12**, 822–834 (2005).
8. Durbin, R., Eddy, S.R., Krogh, A. & Mitchison, G. Biological Sequence Analysis: probabilistic Models of Proteins and Nucleic Acids (Cambridge Univ. Press, Cambridge, 1998).
9. Berman, B.P. *et al.* Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl. Acad. Sci. USA* **99**, 757–762 (2002).
10. Alkema, W.B., Johansson, O., Lagergren, J. & Wasserman, W.W. MSCAN: identification of functional clusters of transcription factor binding sites. *Nucleic Acids Res.* **32**, 169–176 (2004).
11. Sharan, R., Ovcharenko, I., Ben-Hur, A. & Karp, R.M. CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments. *Bioinformatics* **19**, i283–i291 (2003).
12. Donaldson, I.J. *et al.* Genome-wide identification of cis-regulatory sequences controlling blood and endothelial development. *Hum. Mol. Genet.* **14**, 595–601 (2005).
13. Philippakis, A.A., He, F.S. & Bulyk, M.L. Modulefinder: a tool for computational discovery of *cis* regulatory modules. in *Proc. of the Pacific Symp. of Biocomputing* 519–530 (2005).
14. Blanchette, M. *et al.* Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res.* **16**, 656–668 (2006).
15. Zhou, Q. & Wong, W.H. CisModule: De novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc. Natl. Acad. Sci. USA* **101**, 12114–12119 (2004).

16. Rajewsky, N., Vergassola, M., Gaul, U. & Siggia, E.D. Computational detection of genomic cis-regulatory modules applied to body patterning in the early *Drosophila* embryo. *BMC Bioinformatics* **3,** 30 (2002).
17. Sinha, S., van Nimwegen, E. & Siggia, E.D. A probabilistic method to detect regulatory modules. *Bioinformatics*, **19**, i292–i301 (2003).
18. Bailey, T.L. & Noble, W.S. Searching for statistically significant regulatory modules. *Bioinformatics* **19**, 16–25 (2003).
19. Frith, M.C., Li, M.C. & Weng, Z. Cluster-Buster: finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res.* **31**, 3666–3668 (2003).
20. Xie, X. *et al.* Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434**, 338–345 (2005).
21. Wasserman, W.W. & Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* **5**, 276–287 (2004).
22. Sinha, S. *et al.* Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in *Drosophila*. *BMC Bioinformatics* **5**, 129 (2004).
23. Vlieghe, D. *et al.* A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res.* **34**, D95–D97 (2006).
24. Hallikas, O. & Taipale, J. High-throughput assay for determining specificity and affinity of protein-DNA binding interactions. *Nat. Protocols* **1**, 215–222 (2006).
25. Birney, E. *et al.* Ensembl 2006. *Nucleic Acids Res.* **34**, D556–D561 (2006).
26. Kent, W.J. *et al.* The Human Genome Browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
27. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).