W268–W274 *Nucleic Acids Research, 2010, Vol. 38, Web Server issue* *Published online 30 April 2010*
doi:10.1093/nar/gkq330

Phyloscan: locating transcription-regulating binding sites in mixed aligned and unaligned sequence data

Michael J. Palumbo[1] and Lee A. Newberg[1,2,*]

[1]Wadsworth Center, New York State Department of Health, Empire State Plaza, P.O. Box 509, Albany, NY 12201-0509 and [2]Department of Computer Science, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180-3590, USA

Received January 29, 2010; Revised March 26, 2010; Accepted April 17, 2010

ABSTRACT

The transcription of a gene from its DNA template into an mRNA molecule is the first, and most heavily regulated, step in gene expression. Especially in bacteria, regulation is typically achieved via the binding of a transcription factor (protein) or small RNA molecule to the chromosomal region upstream of a regulated gene. The protein or RNA molecule recognizes a short, approximately conserved sequence within a gene's promoter region and, by binding to it, either enhances or represses expression of the nearby gene. Since the sought-for motif (pattern) is short and accommodating to variation, computational approaches that scan for binding sites have trouble distinguishing functional sites from look-alikes. Many computational approaches are unable to find the majority of experimentally verified binding sites without also finding many false positives. *Phyloscan* overcomes this difficulty by exploiting two key features of functional binding sites: (i) these sites are typically more conserved evolutionarily than are non-functional DNA sequences; and (ii) these sites often occur two or more times in the promoter region of a regulated gene. The website is free and open to all users, and there is no login requirement. Address: (http://bayesweb.wadsworth.org/phyloscan/).

INTRODUCTION

With the sequencing of many genomes, we may immediately start asking questions about the genes that are being found. The gene sequences encode proteins and other products, but what do the gene products do and what determines the quantity of expression of a gene product? The answer to the latter question is key to the study of normal and pathological cell function and differentiation; for instance, how does a muscle cell know not to produce proteins used exclusively in skin cells, and how might the regulation go awry?

There are many steps in the creation of a gene product from a gene, starting with transcription, the reading of the DNA template to create an RNA message to be used in subsequent steps. Especially in bacteria, gene regulation is typically achieved via the binding of a transcription factor (protein) or small RNA molecule to the chromosomal region upstream of a regulated gene. The protein or RNA molecule recognizes a sequence within such a promoter region and, by binding to it, either enhances or represses expression of the nearby gene.

With a collection of experimentally verified binding sites for a regulating protein or RNA in hand, or with a motif (pattern)-derived therefrom, it is natural to seek additional genes that are regulated by the same molecule. This computational process is called *scanning* (1–16), and it often includes multi-species data and mathematical models for exploiting phylogenetic/evolutionary relationships (17–20). However, especially because the motif is typically short (6–30 nt in length) and tolerant of variation, the determination as to whether a proposed site is a functional binding site can be difficult. Frequently, attempts to hold the level of false positives low also cause the tools to overlook too many experimentally verified binding sites. Among the purely computational approaches, the phylogeny-based tools have some advantage, because they can exploit conservation across species as suggestive of a functional binding site. *Phyloscan* (21) does particularly well, because it handles phylogenetic relationships whether or not a (multiple) sequence alignment is available, and also because it is able to combine the existence of multiple weak binding sites [a common occurrence (22)] into a statistically strong statement that binding does

*To whom correspondence should be addressed. Tel: +1 518 473 8426; Fax: +1 518 402 4623; Email: phyloscan@wadsworth.org

The authors wish it to be known that, in their opinion, they should be regarded as joint First Authors.

occur somewhere in a promoter region. These traits are advantageous for analyses of large multi-genomic data sets.

The Phyloscan algorithmics paper (21) describes how we use the Neuwald–Green technique (23) to statistically combine evidence from multiple sites within a promoter region, and how we use the Bailey–Gribskov technique (24) to statistically combine evidence across unaligned orthologous sequences. The algorithmics paper also describes the quantitative evaluations of Phyloscan that we have performed, and includes several measures of predictive performance, such as sensitivity, specificity and positive predictive value, as estimated from real and simulated data. Some of the earlier data are reproduced in Figure 1. Note that the '1 clade / 1 site' functionality is similar to that of MONKEY (17), although MONKEY employs techniques to optimize the placement of sequence alignment gaps.

With the new web server, the underlying algorithmics remain unchanged. The new web server permits the user to supply the data to be scanned, where the older server scanned only a specific set of gamma-proteobacterial species data. The new server allows several data formats instead of requiring the use of the FASTA format. Additionally, the new web server provides a tutorial and expanded 'help' information.

The Phyloscan runtime is $O(wL)$, where w is the width of a binding site and L the number of nucleotides in the sequences to be scanned. The constant of proportionality is ~2 μs; Phyloscan scans 2 million nucleotides with a motif model of width 16 in 60 s.

THE INPUTS

For input, Phyloscan requests the information itemized below. Defaults and/or examples are available for each item.

E-mail address

The user can optionally supply an e-mail address. If it is supplied, the user will receive notification when the submitted Phyloscan job has completed. Whether or not an e-mail address is supplied, upon job submission the user will be provided a link to where the results will become available. The user can go to that web page immediately; the page refreshes every 10 s until the results become available.

Phylogenetic tree

Phyloscan exploits phylogenetic relationships among sequences that are (multiply) aligned, by employing nucleotide substitution models: non-functional nucleotides are modeled with HKY85 (25) and binding-site nucleotides are modeled with HB98 (26). To make use of these models, Phyloscan needs a phylogenetic tree relating the species from which the sequences derive. The user should attempt to find an applicable tree in the literature. Alternatively, the user can make an educated guess; Phyloscan will perform well enough if there has been a good-faith effort to give a reasonable tree topology and set of edge lengths.

The phylogenetic tree should be supplied in Newick tree format (also termed New Hampshire tree format); a description for that is available on the Phyloscan help page. The length of each phylogenetic tree edge should be supplied as a non-negative number; it is the average number of substitution events, per nucleotide position, that are expected in neutrally evolving (junk) DNA. For instance, a value of 0.1 for a phylogenetic tree edge means that, within a span of 500 nt positions, we expect an average of 50 nt substitution events to occur, in the time interval separating the ancestral and descendant sequences that are connected by that edge.

Sequences to be scanned

The user selects a file format, and supplies gene promoter (or other) sequence data to be scanned, by pasting them into a text box, or by uploading a file. Each sequence is labeled by the species from which it comes and by the gene (i.e. orthologous gene group) with which it is associated. Sequences can be supplied as aligned or unaligned, and the choice need not be consistent from gene to gene. For instance, suppose that human, chimp and baboon promoter sequences for gene 'abc' are aligned, and the orthologous sequences for mouse and rat are also aligned; when the data for gene 'bcd' is supplied, the promoter sequences from the same species can be grouped differently for alignments, and any of the sequences can be left unaligned to the others. Each supplied sequence should appear exactly once in the input data.

The supplied identifier for a sequence must conform to a specific format. The text before the first '.' must match the name of a species present in the phylogenetic tree. The text after the last '.' must match those sequences that are orthologous to the sequence, whether or not aligned; for example, the sequence upstream of the human 'abc' gene and its orthologous counterparts should be labeled with a shared identifier, such as 'abc.' If an identifier has more than one '.', then the text between the first and last '.' is ignored by Phyloscan. The letters in the nucleotide sequences can be any combination of uppercase and lowercase; Phyloscan ignores the case distinction.

Motif model

The user supplies instances of known binding sites as input to Phyloscan, so that Phyloscan can build a motif model for subsequent scanning. These instances are supplied in a user-specified format; they are pasted into the form or uploaded as a file.

From these data, Phyloscan constructs a product phylogeny model (27), also known as a phylogenetic motif model (28). Phyloscan employs the nucleotide substitution models of HKY85 (25) and HB98 (26) for neutral- and functional-position evolution, respectively.

All supplied binding sites should be unaligned, gapless, and of the same length. Known binding sites can be found in public databases such as JASPAR (29), PAZAR (30) PRODORIC (31), RegTransBase (32) and TRANSFAC (33).

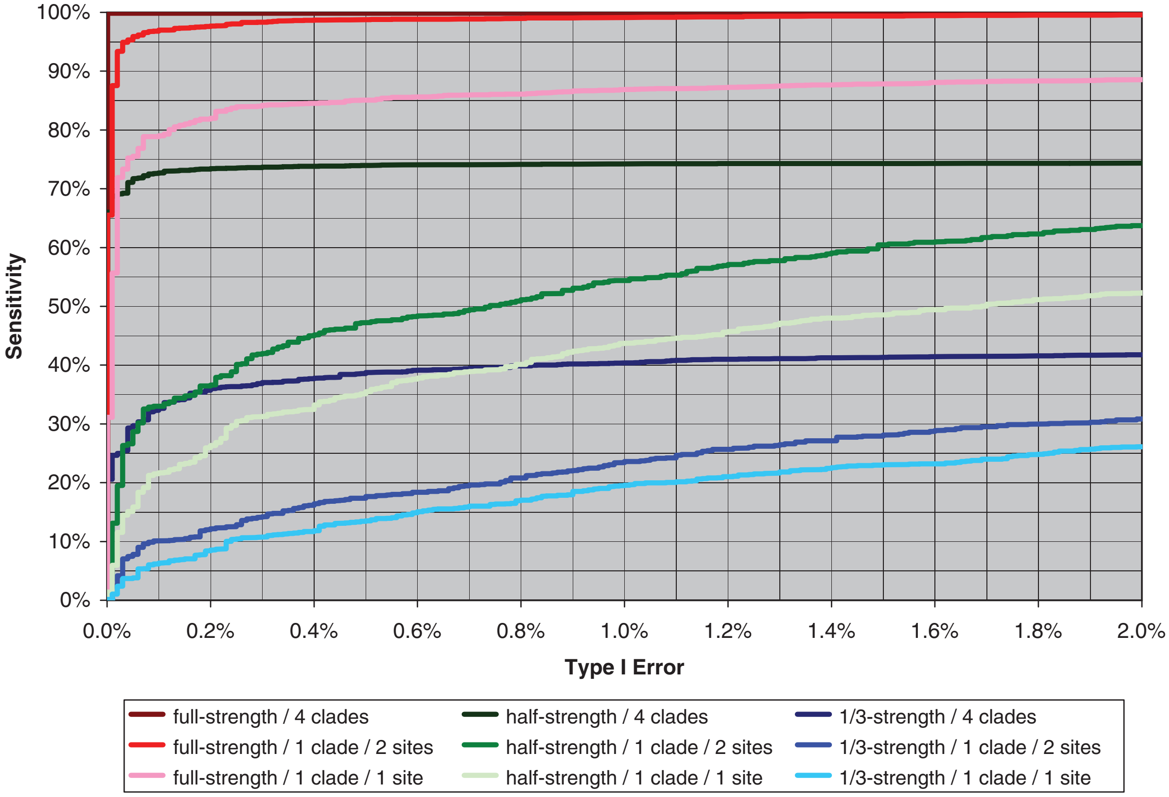


Figure 1. Shown are receiver operating characteristic (ROC) curves for Phyloscan as applied to promoter regions containing a pair of full-strength *Escherichia coli* Crp binding sites, a pair of 1/2-strength sites, and a pair of 1/3-strength sites. The simulated sequence data are for 14 prokaryotic species organized into four clades; the orthologous promoter regions are each 500-nt long and are multiply aligned within each clade, but not between clades. ROC curves are shown for fully enabled Phyloscan, as well as for Phyloscan without the advantage of its multi-clade functionality and for Phyloscan without the advantage of both its multi-clade and its multi-site functionality. (Phyloscan with its multi-clade functionality but without its multi-site functionality is not displayed, because it is nearly indistinguishable from the fully enabled Phyloscan.) A comparison of the '1 clade/2 sites' curves to the '1 clade/1 site' curves shows that there is value in combining evidence from multiple sites within a promoter region, using the Neuwald–Green calculation (23). A comparison of the '4 clades' curves to the '1 clade/2 sites' curves indicates that there is additional value in considering data from multiple clades, using the Bailey–Gribskov calculation (24). For instance, if *p*-value cutoffs are chosen so that the false-positive rate (type I error) is 0.1% (i.e. the specificity is 99.9%), then Phyloscan correctly classifies 99.85% of the full-strength-Crp promoter regions, 72.68% of the 1/2-strength regions and 32.64% of the 1/3-strength regions. The corresponding numbers for '1 clade/2 sites' are 96.98%, 33.01% and 10.11%. The corresponding numbers for '1 clade/1 site' are 79.02%, 21.66% and 6.33%. See the Phyloscan algorithmics paper (21) for further details.

Palindrome flag

The user specifies whether Phyloscan should assume that the supplied known binding sites are palindromes: when a nucleotide sequence (read from 5′ to 3′) is identical to the Watson–Crick complementary sequence to which it would bind in a DNA double helix (also read from 5′ to 3′), the sequence is said to be palindromic.

Many transcription factors are dimeric and recognize a motif that is palindromic; Phyloscan can exploit this common occurrence. Among other features, a check in the palindrome form box permits Phyloscan to skip the reverse scan of each supplied sequence, leading to better statistical significance for the binding sites that are located.

When the user indicates a palindromic model, each binding site supplied as part of the motif model can be supplied in either orientation, but not in both orientations. When the user indicates a non-palindromic model, all of the binding sites supplied for the motif model must have the same orientation, from the perspective of the binding protein or RNA molecule.

Fragmentation mask

Many transcription factors are relatively insensitive to the identity of the nucleotide at some positions within a binding site. For instance, a dimeric transcription factor may bind regardless of the handful of nucleotides that fall between the reverse complement 'half-sites' to which each constituent monomer binds. The user specifies, with an asterisk, which positions are important for binding specificity and, with a period, which positions are ignorable. When in doubt, the user should supply an asterisk for a position.

For example, if the middle six positions of a 22-nt wide binding site are not significant for binding, the supplied fragmentation mask should be

********......********

p-value cutoff

Phyloscan will report a promoter region as being likely to contain one or more binding sites if and only if there is sufficient evidence of the binding sites (i) in the primary species, as considered in isolation and (ii) in the primary species as considered in the context of the remaining orthologous sequences (see below for an explanation of the term primary species). The *p*-value cutoff field sets the cutoff threshold for the primary species considered in isolation; for instance, a cutoff value of 0.05 will instruct Phyloscan to consider only those promoter regions with a *p*-value of 0.05 or better in the primary species. With this cutoff, approximately 1 of 20 promoter regions that do not contain binding sites will be false positives at this stage, and Phyloscan will proceed with the analysis of the promoter region in the context of the promoter region's orthologous sequences. (Such a high interim level of false positives is acceptable because of the further processing that occurs; see *q*-value cutoff below.)

The setting of a low (tight) value for the *p*-value cutoff, e.g. 0.001, will cause Phyloscan to reject promoter regions that do not appear quite good in the primary species, even if they could otherwise be 'rescued' by the existence of high-quality binding sites in the orthologous sequences that are not aligned to the primary species' sequence. Note that a promoter region that passes such a strict cutoff is necessarily of high quality, and frequently such high quality will cause the region to pass the subsequent *q*-value test as well, unless the second test is even more strict. On the other hand, a high (lax) value for the *p*-value cutoff will instruct Phyloscan to not be too concerned with the quality of the binding sites in the primary species; Phyloscan will deem a promoter region to be of high quality if consideration of the primary species and orthologous sequences together so indicates. The default value, 0.05, has been chosen so that Phyloscan will identify (i) those promoter regions that have one or more high-quality binding sites in the primary species and (ii) those promoter regions that have only low-quality binding sites in the primary species but for which the conservation of those sites across the remaining species is significant evidence of the functionality of those sites. However, binding sites that are absent in a promoter region in the primary species, but present in the orthologous sequences, are unlikely to be detected when the cutoff is 0.05 (or lower).

q-value cutoff

The *q*-value cutoff is the mechanism by which Phyloscan balances the trade-off between the number and quality of the promoter regions that it identifies. The *q*-value (also termed the false discovery rate) is the expected ratio of the number of false discoveries in an output data set to the size of the output data set. For example, for a set of 40 promoter regions reported as significant hits by Phyloscan, a *q*-value of 0.05 would indicate that, on average, 2 of those 40 will be false discoveries (under the assumption that the statistical models that are employed perfectly model the underlying biology). This cutoff defaults to 0.001, a conservative value, to account for the fact that the actual biology is more complicated than are the statistical models that we use to analyze it.

Note that *q*-value differs from *p*-value. Each is a fraction with the numerator equal to the number of false positives in an output set. However, for *p*-value the denominator is the expected number of negative cases (i.e. the number of promoters to which the regulatory molecule does not bind); for *q*-value the denominator is the size of the output set.

Rank weights

Much of the strength of Phyloscan arises from its ability to combine the evidence across multiple binding sites within a promoter region. The default weight, 0.9, for the best site indicates to Phyloscan that ~90% of the time, a promoter region with one or more functional binding sites will have at least one strong binding site. The default rank weight, 0.1, for the second-best site indicates to Phyloscan that ~10% of the time, the best site will not be strong, yet the second-best site will be strong enough that the best two sites taken together cause the promoter region to be identified as functional for the transcription factor.

The user must supply one or more rank weights. Each supplied rank weight must be non-negative, and at least one of the rank weights must be positive. If the supplied rank weights do not sum to 1.0, they will be scaled proportionally.

Primary species

Once Phyloscan has accepted the above inputs and has checked that they are reasonable, it will ask the user to select a primary species. This selection influences the algorithm as discussed earlier, in the '*p*-value cutoff' section.

Acknowledgment boxes

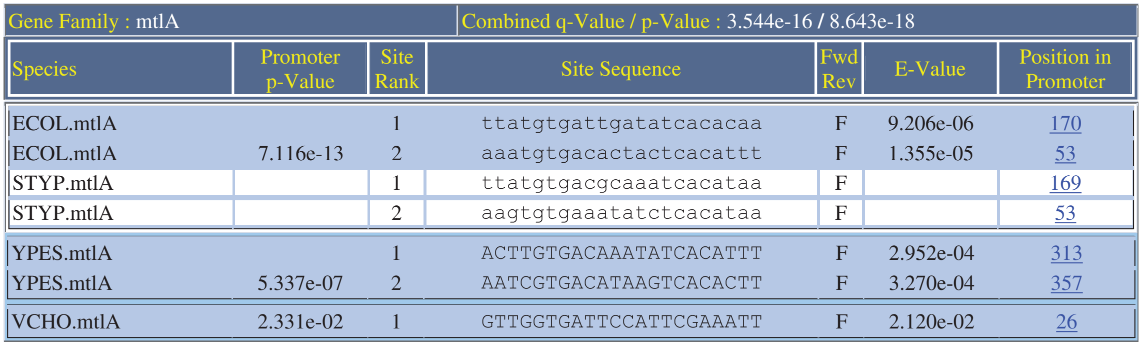
As part of it evaluation of the user-supplied inputs, Phyloscan checks whether any species present in the phylogenetic tree fails to be present in the sequence data and, conversely, whether any species present in the sequence data fails to be present in the phylogenetic tree. If the former event arises, the user is asked to acknowledge that the extra species in the phylogenetic tree will be ignored. If the latter event occurs, the user is asked to acknowledge that the supplied sequences for the extra species will be ignored.

THE OUTPUTS

Figure 2 shows the best result calculated from the example data that is provided by the web site. Here, we describe the fields present in the output.

Gene family

Gene family is the name associated with a gene and its orthologs (if any). It is extracted from the sequences-to-be-scanned input data and is the text following the last '.' in a sequence identifier.



Gene Family : mtlA | Combined q-Value / p-Value : 3.544e-16 / 8.643e-18

Species	Promoter p-Value	Site Rank	Site Sequence	Fwd Rev	E-Value	Position in Promoter
ECOL.mtlA		1	ttatgtgattgatatcacacaa	F	9.206e-06	170
ECOL.mtlA	7.116e-13	2	aaatgtgacactactcacattt	F	1.355e-05	53
STYP.mtlA		1	ttatgtgacgcaaatcacataa	F		169
STYP.mtlA		2	aagtgtgaaatatctcacataa	F		53
YPES.mtlA		1	ACTTGTGACAAATATCACATTT	F	2.952e-04	313
YPES.mtlA	5.337e-07	2	AATCGTGACATAAGTCACACTT	F	3.270e-04	357
VCHO.mtlA	2.331e-02	1	GTTGGTGATTCCATTCGAAATT	F	2.120e-02	26

Figure 2. A run with the example data set provided by our web server, for identifying *Escherichia coli* binding sites for Crp, gives the 'mtlA' gene family as the best result. The combined *q*-value for this gene family, 3.544×10^{-16}, indicates that the user who takes all results of this quality or better (in this case, just the one result) will, 'on average,' find that $<10^{-15}$of the results are false discoveries. The combined *p*-value, 8.643×10^{-18}, indicates that if the user had looked at only the mtlA gene family, and believed the family to be non-functional for Crp binding, then the chance that it would accidentally look this functional for Crp binding is $<10^{-17}$. The combined *p*-value is computed from the promoter *p*-values via the technique of Bailey and Gribskov (24). The promoter *p*-values, 7.116×10^{-13}, 5.337×10^{-7} and 2.331×10^{-2}, arise from the scans of the three user-supplied alignment blocks for mtlA: (i) *E. coli* aligned to *Salmonella enterica* serovar Typhi (*S. typhi*), (ii) *Yersinia pestis* and (iii) *Vibrio cholerae*, respectively. These promoter *p*-values are constructed from the best two, the best two and the best one sites found, respectively, using the technique of Neuwald and Green (23). The best two sites in the *E. coli*–*S. typhi* aligned sequence data have *E*-values of 9.206×10^{-6} and 1.355×10^{-5}; the user can display them in context in, e.g. the *E. coli* sequence, by clicking on the position numbers 170 and 53. The field names in yellow are links to help for these fields.

Combined *q*-value

The combined *q*-value is the proportion of groups of orthologous promoter regions in the Phyloscan output of this quality or better that is expected to be false discoveries. For instance, if $q = 0.05$ for the 40th-best reported promoter region, that result indicates that, on average, 2 among the 40 are false discoveries.

Combined *q*-value is a measure of a promoter region and its orthologous sequences, whether aligned to it or not, when the evidence for all of the sequences and for all of the potential binding sites are considered together. This statistic reflects multiple-testing considerations. Because the statistical model only approximately models the underlying biology, we find that a value ≤ 0.001 to be statistically significant in many circumstances.

Combined *p*-value

The combined *p*-value is the probability that a randomly generated promoter region will accidentally look this good. This statistic does not reflect multiple-testing considerations, in that its computation ignores the number of promoter regions that were scanned. Similar to the combined *q*-value, the combined *p*-value is a measure of a promoter region and its orthologous sequences, whether aligned to it or not, when the evidence for all of the sequences and for all of the potential binding sites are considered together.

Species name

A species name must be associated with each sequence. It will be extracted from the sequences-to-be-scanned input data, as the text preceding the first '.' in the sequence identifier. It is also present in the user-supplied phylogenetic tree.

If, for a gene promoter region, a species' sequence is aligned with one or more orthologous sequences, they will be presented together in a block. The promoter *p*-value (described below) and the binding sites' *E*-values (that are also described below) shown with the first species in the block are statistics applicable to the alignment block.

Promoter *p*-value

Promoter *p*-value is a measure of a single alignment block of a promoter region, when the evidence of all the sequences within the block and all the potential binding sites within the block are considered together. Promoter *p*-value is the probability that a randomly generated alignment block will accidentally look this good. For alignment blocks that contain sequence from the primary species, the promoter *p*-value will be lower than the user-specified *p*-value cutoff.

Site rank

The site rank is the relative strength of a potential binding site found in the sequence data. A value of '1' indicates that it is the strongest site found in a species' sequence data for a promoter region, a value of '2' indicates that it is the second strongest site, and so on.

The number of sites listed will depend upon the user-provided input rank weights and the strengths of the sites. In addition to an evaluation of its *strength*, via the rank weights each site is evaluated as to how *surprising* it is to find a site of this strength at this rank. For example, there are instances for which the discovery that the strongest site has an *E*-value of 0.10 is not unusual, but for which the discovery that the second strongest site has a weaker *E*-value of 0.15 is unusual. All sites that are as strong as or stronger than the most unusual site are listed.

Site sequence

Phyloscan reports the sequence of nucleotides in each potential binding site. Note that these are shown in the forward orientation, even when the site better matches the pattern when read in the reverse-complement sequence.

Sequence orientation

The sequence orientation is set to 'F' when the forward orientation of the potential binding site matches the pattern. It is set to 'R' when the reverse-complement sequence is the match to the pattern. When the pattern is palindromic, an 'F' will always be indicated.

Binding site *E*-value

The binding site *E*-value is similar to the promoter *p*-value, although it does not combine evidence across multiple potential binding sites. The *E*-value for a single potential binding site is the average number of sites in a randomly generated alignment block of this size that are expected to accidentally look this good.

Position in promoter

The location of each potential binding site in the input sequence data is reported. The first position in any input sequence is numbered '1' (rather than '0', as some computer scientists prefer). Gaps are not counted.

Clicking on the number will take the user to a web page that shows the location(s) of the potential binding site(s) graphically.

CONCLUSION

The ability to scan DNA sequence for regulatory binding sites is key to an understanding of gene regulation and its effects on normal and pathological cell function and differentiation. For the first time, our new web server brings together the use of the Bailey–Gribskov technique, for combining mixed aligned and unaligned sequence data, and the Neuwald–Green technique, for statistically combining multiple binding sites' data, into a scan engine that runs on a user's multi-genomic data sets.

FUNDING

National Science Foundation (CCF0914739); Department of Energy (DE-SC000592 to L.A.N.); National Institutes of Health (5K25HG003291 to L.A.N.); Wadsworth Center Bioinformatics Core. Funding for open access charge: National Science Foundation (CCF0914739) and Department of Energy (DE-SC000592).

Conflict of interest statement. None declared.

REFERENCES

1. Hertz,G.Z., Hartzell,G.W. 3rd and Stormo,G.D. (1990) Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci.*, **6**, 81–92.
2. Quandt,K., Frech,K., Karas,H., Wingender,E. and Werner,T. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.*, **23**, 4878–4884.
3. Chen,Q.K., Hertz,G.Z. and Stormo,G.D. (1995) MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Comput. Appl. Biosci.*, **11**, 563–566.
4. Prestridge,D.S. (1996) SIGNAL SCAN 4.0: additional databases and sequence formats. *Comput. Appl. Biosci.*, **12**, 157–160.
5. Berman,B.P., Nibu,Y., Pfeiffer,B.D., Tomancak,P., Celniker,S.E., Levine,M., Rubin,G.M. and Eisen,M.B. (2002) 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.
6. Kim,J.T., Gewehr,J.E. and Martinetz,T. (2004) Binding matrix: a novel approach for binding site recognition. *J. Bioinform. Comput. Biol.*, **2**, 289–307.
7. Loots,G.G. and Ovcharenko,I. (2004) rVISTA 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res.*, **32**, W217–W221.
8. Yellaboina,S., Seshadri,J., Kumar,M.S. and Ranjan,A. (2004) PredictRegulon: a web server for the prediction of the regulatory protein binding sites and operons in prokaryote genomes. *Nucleic Acids Res.*, **32**, W318–320.
9. Osada,R., Zaslavsky,E. and Singh,M. (2004) Comparative analysis of methods for representing and searching for transcription factor binding sites. *Bioinformatics*, **20**, 3516–3525.
10. Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet*, **5**, 276–287.
11. Münch,R., Hiller,K., Grote,A., Scheer,M., Klein,J., Schobert,M. and Jahn,D. (2005) Virtual Footprint and PRODORIC: an integrative framework for regulon prediction in prokaryotes. *Bioinformatics*, **21**, 4187–4189.
12. Su,G., Mao,B. and Wang,J. (2006) A web server for transcription factor binding site prediction. *Bioinformation*, **1**, 156–157.
13. Hiard,S., Marée,R., Colson,S., Hoskisson,P.A., Titgemeyer,F., van Wezel,G.P., Joris,B., Wehenkel,L. and Rigali,S. (2007) PREDetector: a new tool to identify regulatory elements in bacterial genomes. *Biochem. Biophys. Res. Commun.*, **357**, 861–864.
14. Narlikar,L., Gordân,R. and Hartemink,A.J. (2007) A nucleosome-guided map of transcription factor binding sites in yeast. *PLoS Comput. Biol.*, **3**, e215.
15. Whitington,T., Perkins,A.C. and Bailey,T.L. (2009) High-throughput chromatin information enables accurate tissue-specific prediction of transcription factor binding sites. *Nucleic Acids Res.*, **37**, 14–25.
16. Zambelli,F., Pesole,G. and Pavesi,G. (2009) Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. *Nucleic Acids Res.*, **37**, W247–W252.
17. Moses,A.M., Chiang,D.Y., Pollard,D.A., Iyer,V.N. and Eisen,M.B. (2004) MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol.*, **5**, R98.
18. Moses,A.M., Pollard,D.A., Nix,D.A., Iyer,V.N., Li,X.-Y., Biggin,M.D. and Eisen,M.B. (2006) Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS Comput. Biol.*, **2**, e130.
19. GuhaThakurta,D. (2006) Computational identification of transcriptional regulatory elements in DNA sequence. *Nucleic Acids Res.*, **34**, 3585–3598.
20. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.
21. Carmack,C.S., McCue,L.A., Newberg,L.A. and Lawrence,C.E. (2007) PhyloScan: identification of transcription factor binding sites using cross-species evidence. *Algorithms. Mol. Biol.*, **2**, 1.
22. Gertz,J., Siggia,E.D. and Cohen,B.A. (2009) Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature*, **457**, 215–218.

23. Neuwald,A.F. and Green,P. (1994) Detecting patterns in protein sequences. *J. Mol. Biol.*, **239**, 698–712.
24. Bailey,T.L. and Gribskov,M. (1998) Methods and statistics for combining motif match scores. *J. Comput. Biol.*, **5**, 211–221.
25. Hasegawa,M., Kishino,H. and Yano,T.-a. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, **22**, 160–174.
26. Halpern,A.L. and Bruno,W.J. (1998) Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol. Biol. Evol.*, **15**, 910–917.
27. Newberg,L.A., Thompson,W.A., Conlan,S., Smith,T.M., McCue,L.A. and Lawrence,C.E. (2007) A phylogenetic Gibbs sampler that yields centroid solutions for cis regulatory site prediction. *Bioinformatics*, **23**, 1718–1727.
28. Hawkins,J., Grant,C., Noble,W.S. and Bailey,T.L. (2009) Assessing phylogenetic motif models for predicting transcription factor binding sites. *Bioinformatics*, **25**, i339–i347.
29. Portales-Casamar,E., Thongjuea,S., Kwon,A.T., Arenillas,D., Zhao,X., Valen,E., Yusuf,D., Lenhard,B., Wasserman,W.W. and Sandelin,A. (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **38**, D105–D110.
30. Portales-Casamar,E., Arenillas,D., Lim,J., Swanson,M.I., Jiang,S., McCallum,A., Kirov,S. and Wasserman,W.W. (2009) The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences. *Nucleic Acids Res.*, **37**, D54–D60.
31. Münch,R., Hiller,K., Barg,H., Heldt,D., Linz,S., Wingender,E. and Jahn,D. (2003) PRODORIC: prokaryotic database of gene regulation. *Nucleic Acids Res.*, **31**, 266–269.
32. Kazakov,A.E., Cipriano,M.J., Novichkov,P.S., Minovitsky,S., Vinogradov,D.V., Arkin,A., Mironov,A.A., Gelfand,M.S. and Dubchak,I. (2007) RegTransBase—a database of regulatory sequences and interactions in a wide range of prokaryotic genomes. *Nucleic Acids Res.*, **35**, D407–D412.
33. Matys,V., Fricke,E., Geffers,R., Gößling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E. and Kel-Margoulis,O.V. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.