Bioinformation *open access*
Volume 5 Issue 2 **Software**
www.bioinformation.net

xFITOM: a generic GUI tool to search for transcription factor binding sites

Nidhi Bhargava, Ivan Erill*

Department of Biological Sciences, University of Maryland Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, USA; Ivan Erill– E-mail: erill@umbc.edu; *Corresponding Author

Received May 27, 2010; accepted June 9, 2010; published July 6, 2010

Abstract:
Locating transcription factor binding sites in genomic sequences is a key step in deciphering transcription networks. Currently available software for site search is mostly server-based, limiting the range and flexibility of this type of analysis. xFITOM is a fully customizable program for locating binding sites in genomic sequences written in C++. Through an easy-to-use interface, xFITOM allows users an unprecedented degree of flexibility in site search. Among other features, it enables users to define motifs by mixing real sites and IUPAC consensus sequences, to search the annotated sequences of unfinished genomes and to choose among 11 different search algorithms.

Availability: xFITOM is available for download at: http://research.umbc.edu/~erill

Keywords: Transcription factor, binding site, site search, motif prediction, genome analysis

Background:

The discovery and analysis of transcriptional regulatory networks is a key step in elucidating the complex regulatory apparatus of living beings **[1].** Transcriptional regulation is mediated mainly by transcription factors (TF) that bind DNA and can either hinder (repressors) or promote (activators) the formation of an open complex by the RNA-polymerase holoenzyme **[2].** The semi-specific recognition of binding sites by their cognate transcription factors allows implementing computational tools for the discovery and detection of transcription binding motifs and sites **[1].** Motif discovery methods focus on the identification of overrepresented patterns in groups of sequences **[3].** Conversely, site search algorithms take in a motif description and use pattern matching techniques to search for sites on DNA sequences **[4].** Scanning genomic sequences leads to a noisy but informative reconstruction of transcriptional regulatory networks, which can be later validated by in vitro and in vivo methods **[5].**

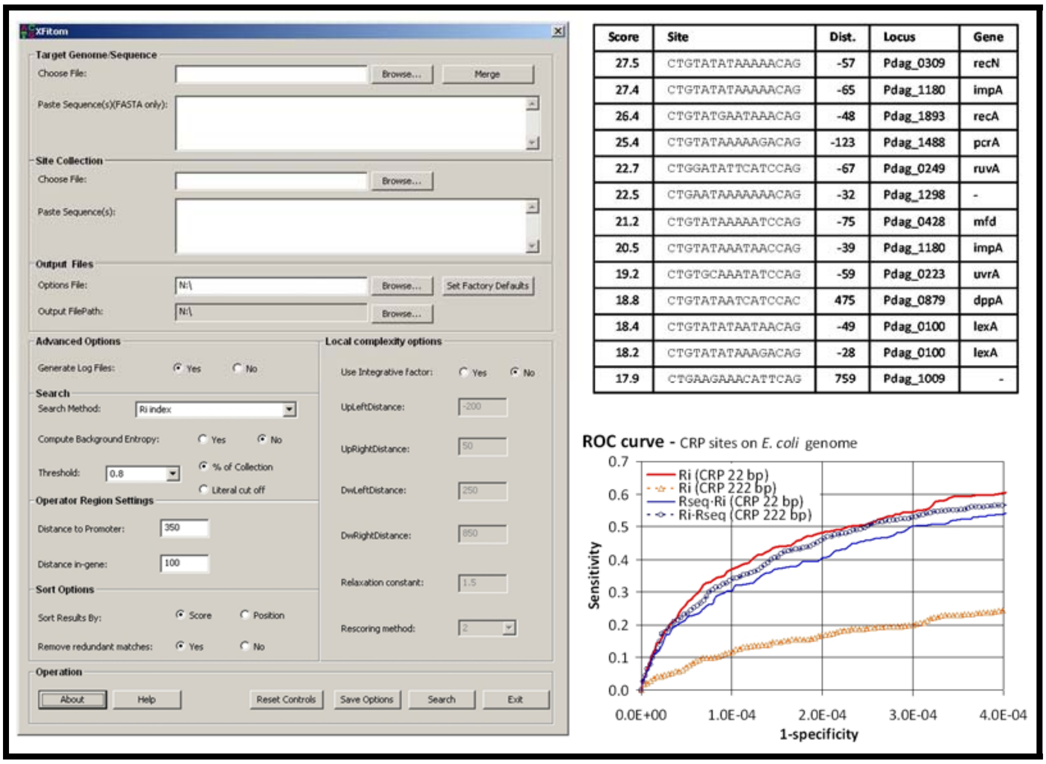
Site search techniques can be roughly divided into two categories. Methods based on based on regular expression syntaxes (regex) use IUPAC codes for degenerate bases and allow for complex motif descriptions including gaps, multiple repeats and variable spacers **[3].** Their flexibility in motif description comes at the cost of reduced base frequency information and a combinatorial increase in complexity for elaborate patterns **[3].** Position-Specific Scoring Matrices (PSSM) approaches implement rigid or semi-rigid feature locators based on the observed frequency of occurrence of each base at each motif position, as derived from a collection of aligned sites **[6].** This matrix representation is commonly referred to as the binding motif or profile. Most PSSM approaches are based on the application of information theory to molecular biology **[4, 7],** which leads to the widespread representation of binding motifs as sequence logos **[8].**

Information theory can be applied to the information process that takes place when a transcription factor binds a site. Protein binding leads to a reduction in uncertainty that is formally defined as the difference in entropy at each motif position **[4] (see supplementary material for equation**). Following this approach, several methods have been devised to assess the fit of a candidate site to a particular binding motif. Non-weighted methods use only the frequency of observed bases at each position in their computation, while weighted methods apply also information on positional conservation. Weighted methods show improved correlation with experimental binding affinities, but are more prone to false positives in genomic searches **[4].** Alternative methods using relative entropy, instead of Shannon entropy, have been devised to take into account biases in genomic base composition **[4].**

Most software solutions developed to implement genome-wide searches of TF-binding sites using PSSM, like Virtual Footprint, are server-based applications linked to server-based databases **[9]** and, in some cases, integrated in web-based motif discovery suites **[3].** A common problem with most web-based applications in bioinformatics is the limitation of sequence sizes in order to contain server load and/or traffic. This is normally addressed by relying on a server database of sequences, but this often limits severely the number and repertoire of target sequences available for analysis. Some solutions rely also on a limited set of pre-compiled matrices, while others restrict entry of motif data to either site collections or IUPAC codes. In addition, each solution relies on a single specific method for site search. This prevents users from choosing the method that they believe most appropriate and precludes the integration of results from different methods. xFITOM is a standalone, easy to use and extremely flexible GUI-based application for site search that integrates several PSSM-based methods in a single tool.

Methodology:

xFITOM requires a collection of aligned sites and a set of target files to be searched. Using the site collection, xFITOM will compute the information content of the motif. It will also pre-process the target sequence and compute its a priori entropy. The program will then start a sequential search of the genome using a sliding window and one of the 11 available scoring methods. xFITOM uses a threshold, an arbitrary cut-off value or a value relative to the distribution of scores in the collection, to define putative binding sites. These are then sorted by genomic position or score. If gene annotation data is available, sites are classified as “operator”, “intragenic” or “intergenic” based on their location with respect to nearby genes. xFITOM can also apply local complexity factors to detect locally enriched motifs **[10].** In this mode, xFITOM computes several mobile score averages and can rescore sites based on three different ratios between local and global score averages.



XFitom

Target Genome/Sequence
Choose File:
Paste Sequence(s)(FASTA only):

Site Collection
Choose File:
Paste Sequence(s):

Output Files
Options File: N:\
Output FilePath: N:\

Advanced Options
Generate Log Files: ● Yes ○ No

Search
Search Method: Ri index
Compute Background Entropy: ○ Yes ● No
Threshold: 0.8 ● % of Collection ○ Literal cut off

Operator Region Settings
Distance to Promoter: 350
Distance in-gene: 100

Sort Options
Sort Results By: ● Score ○ Position
Remove redundant matches: ● Yes ○ No

Local complexity options
Use Integrative factor: ○ Yes ● No
UpLeftDistance: -200
UpRightDistance: 50
DwLeftDistance: 250
DwRightDistance: 850
Relaxation constant: 1.5
Rescoring method: 2

Score	Site	Dist.	Locus	Gene
27.5	CTGTATATAAAAACAG	-57	Pdag_0309	recN
27.4	CTGTATATAAAAACAG	-65	Pdag_1180	impA
26.4	CTGTATGAATAAACAG	-48	Pdag_1893	recA
25.4	CTGTATAAAAAGACAG	-123	Pdag_1488	pcrA
22.7	CTGGATATTCATCCAG	-67	Pdag_0249	ruvA
22.5	CTGAATAAAAAAACAG	-32	Pdag_1298	-
21.2	CTGTATAAAAATCCAG	-75	Pdag_0428	mfd
20.5	CTGTATAAATAACCAG	-39	Pdag_1180	impA
19.2	CTGTGCAAATATCCAG	-59	Pdag_0223	uvrA
18.8	CTGTATAATCATCCAC	475	Pdag_0879	dppA
18.4	CTGTATATAATAACAG	-49	Pdag_0100	lexA
18.2	CTGTATATAAAGACAG	-28	Pdag_0100	lexA
17.9	CTGAAGAAACATTCAG	759	Pdag_1009	-

Figure 1: [Left panel] GUI interface of xFITOM showing the input section and the advanced and local complexity options sections. [Top-right panel] Condensed results from an annotated search of the *Pasteurella dagmatis* ATCC_43325 unfinished genome sequence using a collection of LexA-binding sites from *Escherichia coli* **[4].** For clarity, gene locus accession numbers have been abbreviated to Pdag_XXXX from the original HMPREF0621_XXXX format. Searching merged GenBank files can provide useful information on the composition regulatory networks for partly sequenced organisms. Here it shows the usual components of the SOS network in the γ-Proteobacteria to be present in *P. dagmatis* **[11].** [Bottom-right panel] ROC curve showing a comparison of search efficiencies when looking for CRP sites in the *E. coli* genome with different methods. Previous results **[4]** indicate that non-weighted methods like R_i outperform weighted ones (like $R_i{\cdot}R_{sequence}$), but this does not hold true when the length of sites in a collection extends beyond the primary conserved region.

Implementation:

xFITOM has been developed integrally in C++ as a standalone application for Microsoft Windows operating systems using the Microsoft Foundation Classes. The program operates on two main input files: a collection of aligned sites and a collection of target sequences. Aligned sites can be entered in FASTA and raw text format, with degenerate IUPAC characters allowing for an arbitrary degree of flexibility in the definition of the binding motif. Target sequences can be entered in FASTA and GenBank formats, with gene annotation data extracted automatically from the latter. Multiple sequences can be processed directly in FASTA format, while GenBank files can be merged into a compound GenBank file for analysis. This allows analysis of annotated unfinished genomes, giving immediate access to newly released genomic data. All xFITOM parameters can be specified by the user through the GUI (**Figure 1**). Basic parameters, like the search method, detection threshold and site-to-gene distances are defined in the left panel, while local complexity options are located on the right panel. The program generates comma-separated value output files to allow easy post-processing of results with spreadsheet software. Due to its standalone nature and configurability, xFITOM allows users to carry out useful non-standard analyses, such comparisons among different methods (**Figure 1**), affinity ranking of collection sites or annotated searches of unfinished genome assemblies (**Figure 1**).

Acknowledgements

This research was supported partly by the Summer Faculty Fellowship program at UMBC.

References:

[1] MM Babu, *Biochem Soc Trans*, **36**(4): 758 (2008). [PMID:18631154]

[2] M Ptashne, *Trends Biochem Sci*, **30**(6): 275 (2005). [PMID:15950866]

[3] J Mrazek, *Brief Bioinform*, **10**(5): 525 (2009) [PMID: 19553402]

[4] I Erill & MC O'Neill, *BMC Bioinformatics*, **10**(1): 57 (2009). [PMID:]

[5] Erill, I *et al. Nucleic Acids Res*, **32**(22): 6617 (2004). [PMID: 15604457]

[6] S Hannenhalli, *Bioinformatics,* **24**(11):1325 (2008). [PMID: 18426806]

[7] TD Schneider *et al. J Mol Biol*, **188**(3): 415 (1986) [PMID: 3525846]

[8] TD Schneider & RM Schneider, *Nucleic Acids Res,* **18**(20): 6097 (1990) [PMID: 2172928]

[9] R Munch *et al. Bioinformatics*, **21**(22):4187 (2005) [PMID: 16109747]

[10] SE Halford, & JF Marko, *Nucleic Acids Res*, **32**(10): 3040 (2004) [PMID: 15178741]

[11] I Erill *et al. Bioinformatics,* **19**(17): 2225 (2003) [PMID:14630651]

Edited by P. Kangueane

Citation: Bhargava *et al.* Bioinformation 5(2): 49-51 (2010)

Bioinformation Volume 5 *open access*
www.bioinformation.net Issue 2 **Software**

Supplementry material:

The theoretical framework of information theory can be applied to the information process that takes place when a transcription factor binds its cognate sites. Given a DNA fragment of unknown sequence, information theory defines a priori and a posteriori entropies as average measures of our uncertainty over the base occupying each position of the DNA fragment before and after we observe binding of a particular transcription factor to it **[4, 7].** The reduction in uncertainty brought about by the binding event can be interpreted as an information gain and is formally defined as the difference between a priori (H_{before}) and a posteriori (H_{after}) entropies at each position:

$$R_{sequence}(l) = H_{before}(l) - H_{after}(l) = \left[-\sum_{S \in \Omega} \left(f(S) \cdot (\log_2 \left(f(S) \right) \right) \right] - \left[-\sum_{S_l \in \Omega} \left(p(S_l) \cdot \log_2 \left(p(S_l) \right) \right) \right] \qquad (1)$$

where S corresponds to each of the four possible DNA bases, *f(S)* represents the relative frequency of *S* in the genome sequence and $p(S_l)$ is the frequency of base S_l at position l in a set of know binding sites. Intuitively, Equation 1 conveys the notion that our best initial guess over what base occupies any given motif position must rely on background genomic frequencies *f(S)*, while the knowledge of binding allows us to refine our guess based on frequencies [$p(S_l)$] derived from a collection of known sites.