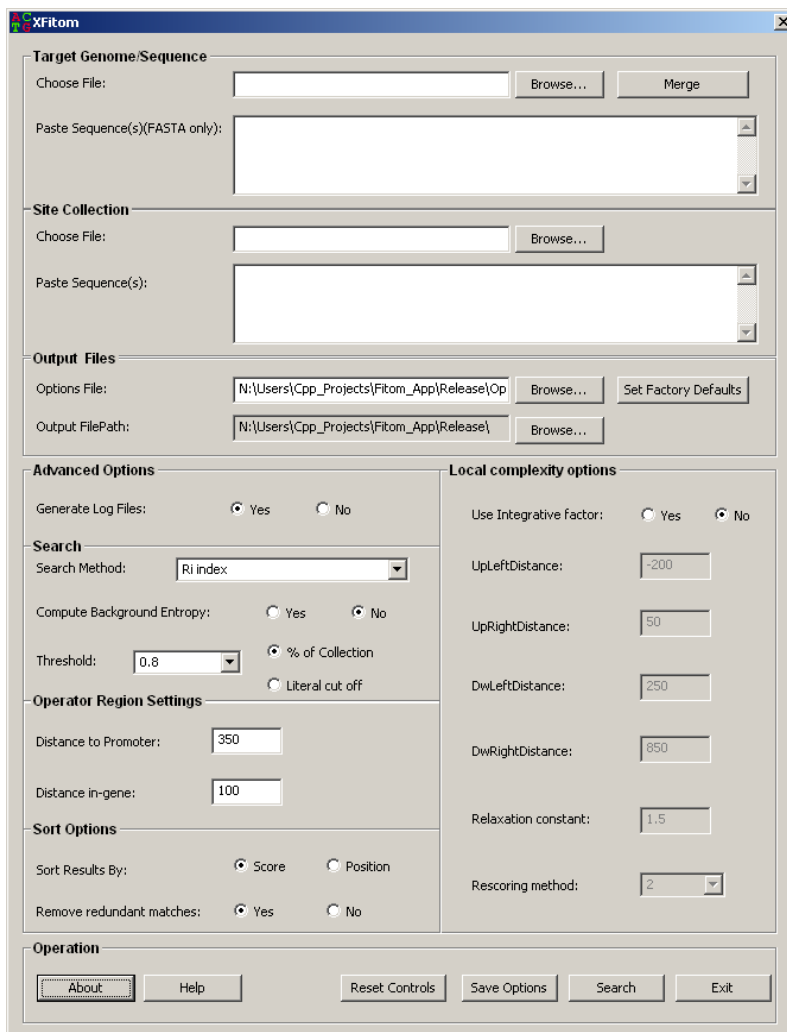


## Introduction

**xFITOM** is a Microsoft Windows graphical version of **FITOM**, a computer program for the detection of binding sites in DNA or RNA sequences. **xFITOM** implements several methods described in the literature to compute an approximation of binding affinity for a particular site based on a collection of binding sequences provided by the user. Using these methods, **xFITOM** scans a sequence file looking for putative binding sites across the DNA/RNA sequence in both strands, and filters the results according to a user-specified threshold. If sequence annotation is provided in the sequence file, **xFITOM** will also link the identified sites with annotated genes and it will infer their role from their location in the vicinity of genes.



**Figure 1:** Graphical User Interface of **xFITOM**

## Credits

Original **FITOM** code by Ivan Erill. **xFITOM** GUI development by Nidhi Bhargava & Ivan Erill. @ Ivan Erill 2010.

If using for research, please cite: Erill, I; O'Neill, M.C. 'A reexamination of information theory-based methods for DNA-binding site identification' BMC Bioinformatics. 2009 Feb 11;10(1):57.

## Getting started

**xFITOM** predecessor, **FITOM**, is a command-line argument based program, meaning that it is run from a DOS prompt. **xFITOM** includes a Graphical User Interface (GUI) to select the required files and to set all the necessary options. In addition, **xFITOM** provides new functionality, like the ability to work with partially completed annotated genome sequences.

### Main operation files in **xFITOM**:

Like its predecessor, **FITOM**, **xFITOM** operates with three main files: a file containing the sequence to be searched (*sequence file*), a file containing a list of binding sites (*collection file*) and a file specifying the program options (*options file*).

#### *The sequence file: (Genome/Sequence file)*

The sequence file (Sequence\_file.ext) is the file containing the sequence or sequences the user wants to scan. These files can be in two possible formats: FASTA and GenBank, carrying the respective extensions. **xFITOM** only accept extensions: FAS/FNA for FASTA files and GBK/GB for GenBank files.

Only once sequence per file is accepted for GenBank format in **FITOM**, while multiple sequences per file are allowed in FASTA format. **xFITOM** provides a "Merge" option (see below) to integrate multiple-sequence GenBank files into a single file for processing.

#### *The collection file: (Site/Collection file)*

The collection file (Collection\_file.ext) is the file containing the collection of known binding sites that the user provides the program with in order to construct its model of binding site, or motif. Collection files can be either bare site files (plain text with aligned sites on consecutive lines) or FASTA files, in which each site line is preceded by an identification line beginning with '>'. Accepted extensions are: FAS/FNA for FASTA files and TXT for bare site files.

#### *The options file: (Options file)*

All options can be set in the options file. The options file (Options\_file.ext) is a simple text file containing all the advanced options. **xFITOM** allows setting the options through its GUI and to save them into a user-defined options file. If the options file is not present then options can be set to default values and automatically saved to an options file. By default, **xFITOM** expects and loads at start-up an Options.txt file in its operating folder. If this file is not found, **xFITOM** reverts to default options.

## Main functionalities:

The xFITOM GUI provides three main functionalities, dealing with different aspects of the program operation.

### *Input File processing*

- *Load sequence file*

It allows user to load file containing the sequence or sequences the user wants to scan. The file can be in two possible formats: FASTA or GenBank. By default user can select only once sequence per file in GenBank format and multiple sequences per file in FASTA format.

xFITOM allows user to use sequence file with multiple Genbank entries. The “Merge” button integrate multiple-sequence GenBank file into a new unified file with all the multiple Genbank entries (and their annotations) properly concatenated.

- *Load collection file*

It allows user to select a file containing a list of known binding sites, which are used to constructor model of binding site or motif. The file can be in two formats: FAS/FNA for FASTA files or TXT for bare site files.

- *Set output path*

It allows user to select the path for saving the output files. If log option is set all the log files generated by xFITOM will be saved in the selected output file path. By default the application path is selected as output file path.

### *Options setting*

- *Load Options file*

The user can select the Options file containing all the advanced options (described below). By default when xFITOM is initiated it sets the options in GUI from Options file available in application folder, if Options file is not available xFITOM sets the default factory values in GUI.

- *Set options*

The main part of xFITOM GUI is devoted to controls on the different advanced options offered by the program and described in detail below.

- *Set Factory Defaults*

The “Set Factory Defaults” button allows the user to set the default options in GUI and automatically save all the options setting to Options.txt file.

- *Save Options file*

The “Save Option File” button allows user to save the current option setting to user defined options file.

## Basic operation

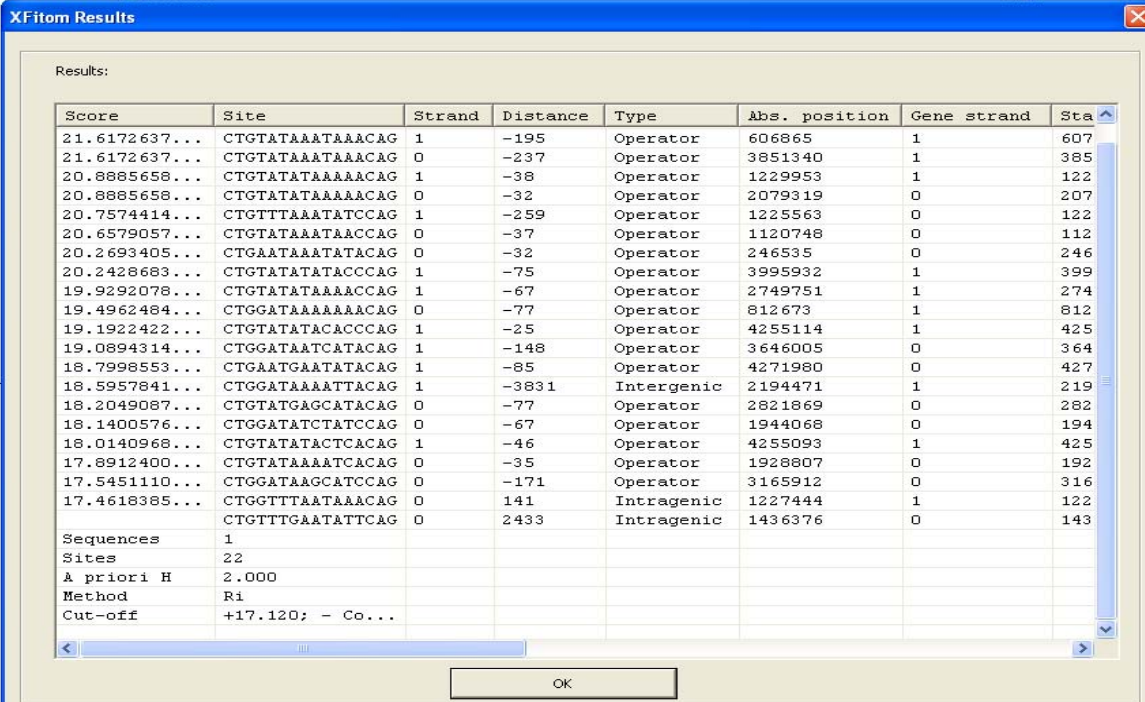
- “Search” button

When the user clicks on “Search” button, xF<sub>i</sub>TOM will proceed to scan the provided sequence looking for putative binding sites across the DNA/RNA sequence in both strands (see main operation description below). After completing the scan, xF<sub>i</sub>TOM will filter the results according to a user-specified threshold. The results will be displayed by xF<sub>i</sub>TOM in an emergent results window and will be saved to a .CSV file in the specified output folder. The CSV (comma-separated value) file can be imported directly by MsExcel or any other spreadsheet program for additional processing and analysis.

- Display results

After the search is complete, xF<sub>i</sub>TOM will display the result in Results window (as shown below). The results contain the identified binding sites, together with their score, position and strand. For more than one sequence, the results file will be divided into consecutive sequences separated by a sequence-name row. At the end of the file, the total number of sequences and sites, as well as the a priori entropy, selected method and threshold are displayed. In case gene information is available, the results file will also contain a site tag, the distance from site to gene start, and basic gene information.

xF<sub>i</sub>TOM will save the results of the analysis into a comma-separated value (CSV) file (Sequence\_file\_name)\_(Collection\_file\_name)\_(method)\_(threshold).csv that can be opened directly with spreadsheet software such as Ms Excel. The result file is saved at the output file path selected by the user.



The screenshot shows a window titled "XFitom Results" with a table of results. The table has columns for Score, Site, Strand, Distance, Type, Abs. position, Gene strand, and Sta. Below the table, there is a summary section with labels like Sequences, Sites, A priori H, Method, and Cut-off.

Score	Site	Strand	Distance	Type	Abs. position	Gene strand	Sta
21.6172637...	CTGTATAAAATAAACAG	1	-195	Operator	606865	1	607
21.6172637...	CTGTATAAAATAAACAG	0	-237	Operator	3851340	1	385
20.8885658...	CTGTATATAAAAACAG	1	-38	Operator	1229953	1	122
20.8885658...	CTGTATATAAAAACAG	0	-32	Operator	2079319	0	207
20.7574414...	CTGTTTAAATATCCAG	1	-259	Operator	1225563	0	122
20.6579057...	CTGTATAAAATAACCAG	0	-37	Operator	1120748	0	112
20.2693405...	CTGAATAAATATACAG	0	-32	Operator	246535	0	246
20.2428683...	CTGTATATATACCCAG	1	-75	Operator	3995932	1	399
19.9292078...	CTGTATATAAAAACAG	1	-67	Operator	2749751	1	274
19.4962484...	CTGGATAAAAACAG	0	-77	Operator	812673	1	812
19.1922422...	CTGTATATACACCCAG	1	-25	Operator	4255114	1	425
19.0894314...	CTGGATAATCATAACAG	1	-148	Operator	3646005	0	364
18.7998553...	CTGAATGAATATACAG	1	-85	Operator	4271980	0	427
18.5957841...	CTGGATAAAATTACAG	1	-3831	Intergenic	2194471	1	219
18.2049087...	CTGTATGAGCATACAG	0	-77	Operator	2821869	0	282
18.1400576...	CTGGATATCTATCCAG	0	-67	Operator	1944068	0	194
18.0140968...	CTGTATATACTCACAG	1	-46	Operator	4255093	1	425
17.8912400...	CTGTATAAAATCACAG	0	-35	Operator	1928807	0	192
17.5451110...	CTGGATAAGCATCCAG	0	-171	Operator	3165912	0	316
17.4618385...	CTGGTTTAAATAAACAG	0	141	Intragenic	1227444	1	122
	CTGTTTGAATATTCAG	0	2433	Intragenic	1436376	0	143
Sequences	1						
Sites	22						
A priori H	2.000						
Method	Ri						
Cut-off	+17.120; - Co...						

Figure 2: Sample Results Window

## Main operation

Before going into the description of the advanced options, it is interesting to describe the main modus operandi of the program. As mentioned above, **xFTOM** starts by loading the provided sequence and collection files.

### Position-specific weight matrix and information content

The collection file is then used to compute the motif position-specific frequency matrix (PSFM). This matrix is a matrix of the relative frequencies of each nucleotide at each position in the motif:

	1	2	3	4	5	6
A	0.031	0.055	0.650	0.349	0.309	0.007
C	0.928	0.015	0.015	0.071	0.158	0.007
G	0.007	0.206	0.166	0.031	0.079	0.976
T	0.031	0.722	0.166	0.547	0.452	0.007
Consensus	C	T	A	T	T	G

From the positions specific frequency matrix, the information content ( $R_{sequence}$ ) or redundancy index of the motif can be computed, according to the following formula:

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

$f(S)$  – frequency of base  $S$  in the genome

$p(S)$  – frequency of base  $S$  in the motif PSWM

(**xFTOM** adds  $10^{-100}$  to each motif frequency to avoid  $0 \cdot \log(0)$  terms for bases not represented in the collection)

$H_{before}$  – a priori entropy

$H_{after}$  – entropy after binding

as described by Schneider *et al.* (Schneider, Stormo *et al.* 1986) and based on the assumption of positional independency among the different positions of a binding site.

The information content of a motif tells us about the reduction in uncertainty we experience once we know that a protein (or other element) binds to a sequence (Schneider, Stormo *et al.* 1986; Erill and O'Neill 2009). Prior to binding, our uncertainty about what bases occupy the different positions of a sequence is maximal, and dictated by the base composition of the genome. Once we know that the protein associated with the provided motif binds that sequence, however, we have much less uncertainty about what bases occupy the different positions. We still have uncertainty, because protein binding is a noisy issue, but we have decreased our uncertainty and, thus, we can say we have gained information. Conversely, seen from the point of view of a genome, the information content can also be seen as the loss of entropy at certain regions in the genome, from an initial random state to a state of fixation of conserved binding sites. Thus, motif information content can also be as an index of the level of redundancy ( $RI$ ) in the different positions of the motif (O'Neill 1998).

Even though without a complete theoretical justification, a different index termed relative entropy ( $RE$ ) has been proposed to substitute the  $RI$  in cases of heavily skewed genomes:

$$RE(l) = \sum_{S \in \Omega} \left( p(S_l) \cdot \log_2 \left( \frac{p(S_l)}{f(S)} \right) \right)$$

Relative entropy (Schneider, Stormo et al. 1986; Erill and O'Neill 2009) is also computed by **xFrToM** and can be used in different ranking methods.

### Ranking methods

$R_{sequence}$  tells us how much information our motif conveys, but it does not provide answers to how well a particular sequence fits in the motif profile, which is what is required to scan for and rank putative binding sites.

Several ranking methods have been proposed with diverse degrees of theoretical justification. **xFrToM** provides two basic scoring methods that can be used to rank putative binding sites. The sequence information content ( $R_i$ ) (Schneider 1997) is a method derived from the information content ( $R_{sequence}$ ) formula that scores each position of a particular site ( $j$ ) based on ratio of frequency in the motif with respect to genomic frequency for the particular base observed in the site:

$$R_i = I^j(l) = \left[ - \sum_{S \in \Omega} (f(S) \cdot (\log_2(f(S)))) \right] - \left[ - \log_2 \left( \left( \frac{p(S_l^j) + 1/N}{1 + 4/N} \right) \right) \right]$$

$(1/N)/(1+4/N)$  is the zero-frequency correction following Laplace Law of Succession to estimate the frequency of a base present in a sequence and not present in the collection.  $N$  is the number of sequences in the collection.

Another proposed method is the Berg-von Hippel Heterology Index ( $HI$ ), based on the relative frequency of the observed base with respect to the dominant (consensus) base at each position (Berg and von Hippel 1987):

$$BvH = HI = \sum_{l=1}^L \ln \left( \frac{P(l_{cons}) + 1/N}{P(l_{obs}) + 1/N} \right)$$

$1/N$  is again a correction factor for bases with zero frequencies in the collection

It has been shown that both methods compute, effectively, the same index (Erill and O'Neill 2009). As a result of their formulation, both ranking methods discard information from the other motif base frequencies. As explained in (O'Neill 2003), this can lead to erroneous scoring, where the same score may be given to little or heavily conserved positions since information about the redundancy of each position (the information from the rest of bases at that position discarded by both methods) is not used. To correct this, O'Neill proposed averaging this kind of methods with the know redundancy index of the collection (O'Neill 1989), so that the final score was given by:

$$R_{sequence} \cdot BvH = \sum_{l=1}^L R_{sequence}(l) \cdot BvH(l) \quad R_{sequence} \cdot R_i = \sum_{l=1}^L R_{sequence}(l) \cdot R_i^j(l)$$

An even more interesting approach also proposed by O'Neill was to solve some of the different problems of the methods described above. The differential  $R_{sequence}$  ( $R'_{sequence}$ ) is a method to score putative binding sites based on the successive evaluation of the  $R_{sequence}$  before ( $R_{sequence}^-$ ) and after ( $R_{sequence}^+$ ) assuming that the site is a member of the collection.

$$R_{sequence}'(l) = R_{sequence}^-(l) \cdot (R_{sequence}^+(l) - R_{sequence}^-(l))$$

Computing then the difference between both  $R_{sequence}$  values provides a measure of how well does the putative site sit within the collection. If the site does not agree with the motif profile generated by the collection, the expanded  $R_{sequence}$  ( $R_{sequence}^+$ ) will decrease with respect to  $R_{sequence}^-$  and the difference will be negative. Conversely, if the site agrees well with the motif profile,  $R_{sequence}^+$  will increase and the difference will be positive.

#### *Cut-off and tagging*

A priori, every subsequence of the same length as the motif is a putative binding site. Therefore, as **xFiToM** scans the genomic sequence, it must discard false positive sites and save only what it considers true positives. This is done using one of the ranking methods described above and a threshold that, again, is provided by the user. Since the specific cut-off value depends on the method, it is often more convenient to specify it using a relative threshold. This value, in the [0-1] range, specifies how many sites from the original collection should be above the cut-off value. Therefore, if the relative threshold is 0.5, a cut-off value will be chosen by **xFiToM** so that, assuming a normal distribution, only 50% of the collection sites is above the threshold.

If available in the sequence GenBank file, **xFiToM** will also gather information on the location of genes in the genome. Using this information and the position of identified binding sites, **xFiToM** will correlate the results in order to link sites to genes. Following user-defined limits, **xFiToM** will assign different tags to sites, labeling them as *operator* (within limits), *intragenic* (within coding region) or *intergenic* (between genes)

#### *Results*

**xFiToM** will save the results of the analysis into a comma-separated value (CSV) file (Sequence\_file\_name)\_(Collection\_file\_name)\_(method)\_(threshold).csv that can be opened directly with spreadsheet software such as Ms Excel. The results file contains the identified binding sites, together with their score, position and strand. In case gene information is available, the results file will also contain a site tag, the distance from site to gene start, and basic gene information.

For more than one sequence, the results file will be divided into consecutive sequences separated by a sequence-name row. At the end of the file, the total number of sequences and sites, as well as the a priori entropy, selected method and threshold are displayed.

#### *Local complexity*

**xF<sub>1</sub>T<sub>0</sub>M** allows taking into account local complexity (in the form of signal overrepresentation) in the detection of binding sites. The idea, developed initially for bacterial promoter detection, is based on the proposed hypothesis that RNA-polymerase (and other DNA-binding proteins) may use weak binding sites upstream of true promoters to improve its promoter-seek dynamics, as 3D diffusion appears to be too limiting to account for the efficiency of RNA-polymerase in transcription (Berg, Winter et al. 1981; Ricchetti, Metzger et al. 1988; Halford and Marko 2004). **xF<sub>1</sub>T<sub>0</sub>M** computes the mean score (ranking) of sites in intervals both upstream and downstream of the site under evaluation, so that the current site score could be corrected according to a local complexity measure that took into account the presence of nearby pseudo-sites (integrative correction).

In integrative mode **xF<sub>1</sub>T<sub>0</sub>M** proceeds normally and scans the sequence in both strands, but it does so using a look-ahead method. This means that **xF<sub>1</sub>T<sub>0</sub>M** will pre-compute the mean score of upstream and downstream regions for the first sequence position (assuming circular DNA). This is called the pre-run. Once this initial means have been computed, **xF<sub>1</sub>T<sub>0</sub>M** scans the sequence and adds the new score to the current means. The site under evaluation then changes to the middle site in the mean-computing interval. A relaxed threshold is used to allow a substantially larger number of candidate sites, which are saved together with the mean values of their surroundings. Once the sequence has been scanned, selected sites are re-evaluated by multiplying their score with a correction factor derived from a ratio between means (e.g. upstream/downstream mean).



## Advanced options:

This section described in detail **xFIToM** advanced options. Advanced options are stored in an Options File that, by default, is named `Options.txt`. In this file, options are set simply by assigning each parameter with a positive integer value in a single line preceded by a \$ sign. **xFIToM** allows direct setting of these parameters through the application GUI, but the option file is still directly modifiable by the user. The following is a description of **xFIToM** options following their natural order in the Options File:

### 1 - Log option

The log option simply determines whether the program should log partial results during operation. Default Log option is *yes*. If the log option is set, **xFIToM** will generate the following files:

- `(Sequence_file_name)_seqs_log.txt`  
contains the read DNA sequences; can be used to extract the DNA sequence in FASTA format from a GenBank source.
- `(Sequence_file_name)_site_collection_log.txt`  
contains the read site collection in FASTA format
- `(Sequence_file_name)_genes_log.csv`  
contains the list of genes read from an annotated GenBank file
- `(Sequence_file_name)_(Collection_file_name)_freq_table.csv`  
contains the frequency table constructed for the read site collection, and the computed positional information content
- `(Sequence_file_name)_(Collection_file_name)_motif_score_log.csv`  
contains the list of read sites and the score associated to each of them, as well as the cutoff value
- `(Sequence_file_name).log`  
displays incidences on the program run

**xFIToM** will also generate a `Fitom.log` file if any problem is encountered while reading files and the program must stop.

### 2- Method option

The method option allows the user to choose between the different methods described above. By default Method option is set to  $R_i$  index, (information content of an individual DNA sequence). Following are the different methods available:

1. Information content of a individual DNA sequence ( $R_i$  index)
2. Berg & von Hippel Heterology Index (HI)
3.  $R_{sequence}$  averaged Berg & von Hippel Heterology Index  $R_{sequence} \cdot BvH$
4.  $R_{sequence}$  averaged differential  $R_{sequence}$  ( $R_{sequence}'$ )
5.  $R_{sequence}$  averaged individual sequence information content ( $R_{sequence} \cdot R_i$ )
6. PredictRegulon Index  
The index described in (Yellaboina, Seshadri et al. 2004) for the PredictRegulon server.
7. RE averaged individual sequence information content ( $RE \cdot R_i$ )  
The same as  $R_{sequence} \cdot R_i$  but using relative entropy (RE) instead of information content ( $R_{sequence}$ ) as the averaging factor.
8. Differential RE ( $RE'$ )  
The same as  $R_{sequence}'$  but using relative entropy (RE) instead of information content ( $R_{sequence}$ ).
9.  $I_{seq}$   
Derives from RE in the same way as  $R_i$  derives from  $R_{sequence}$ . Described in (Hertz, Hartzell et al. 1990).
10. FitomHI ( $RE'$ )  
A modification of Berg & von Hippel heterology index (HI) to truly account for the divergence between consensus and observed bases.
11. Differential  $R_{sequence}$  (Non-weighted  $R_{sequence}'$ )  
A modification of the differential  $R_{sequence}$  method to convert it into a non-weighted method.

### 3 - Background entropy option

In computing the background entropy  $H_{before}$ , several authors have proposed assuming equiprobability [ $H_{before}=2$  bits], irrespective of the genome composition, on the argument that a protein does not know about genome composition in skewed genomes and, hence, its a priori uncertainty should be assumed to be maximal (Schneider, Stormo et al. 1986).

Even though the argument can be disputed, mainly because the protein will have also evolved in the skewed genome, xFIToM allows the user to specify whether to use a fixed 2 bits background entropy (*No*) or to derive it from genome composition (*Yes*). In case of a FASTA file with multiple sequences, xFIToM will assume that  $H_{before}$  is 2 bits, regardless of this option's setting.

### 4 - Threshold

The value introduced in the threshold option can specify two different parameters, depending on the setting of option 9 (*Literal cut-off*). If option 9 is set to *% of collection*, xFIToM will use the value specified in this option (0-1) as a threshold relative to the collection of sites provided by the user. xFIToM will assume that the provided collection of binding sites has a normal distribution of scores and will determine the (method-dependent) cut-off that selects the percentage of sites from the collection indicated by the user through this parameter. If option 9 is set to *literal cut-off*, xFIToM will use the value provided in this option as the cut-off for the chosen method.

The default value is 0.8 (for relative threshold); i.e. threshold at 80% of collection sites.

### 5 - Distance to promoter option

If available, identified sites are tagged with gene information. An important parameter in this tagging, in order to further filter the results provided by xFIToM, is to determine whether a particular site may or may not be an operator (i.e. a site involved in promoter regulation). Since known prokaryote operator sequences fall within a range of the translational start point, xFIToM relies on two user-provided parameters: the distance to promoter and the distance in-gene presets. The first makes reference to the maximum distance a site can be upstream of the gene translational start point (TLS) in order to tag it as *operator*.

Default value is 350 bp; i.e. 350 bp max distance upstream of TLS for operator site.

### 6 - Distance in-gene option

This second distance parameter (see above) makes reference to the maximum distance a site can be downstream of the gene translational start point in order to tag it as *operator*.

Default value is 100 bp; i.e. 100 bp max distance downstream of TLS for operator.

### 7 - Sorting method option

Results can be sorted according to two main parameters of identified binding sites: their position within the genome (*Position*) or their ranking (*Score*). Results can later be resorted in Excel according to user needs.

Default *Sort* option is set to *Score*; i.e. sort results by ranking.

### 8 - Remove redundant matches option

Since the sequence is scanned in both strands, for palindromic motifs a same site will be located in both scans. In some cases, if the site is particularly good, both versions will be picked up as positive results, leading to duplicity of results. The remove redundant matches option allows the user to remove (*Yes*) or not (*No*) these duplicate sites. Duplicate removal proceeds based on a best-ranking policy: among the two versions of the same site, only the better scoring one will be preserved.

Default *Remove redundant matches* option is *Yes*; redundant matches will be removed.

### 9 - Literal cut-off or % of Collection option

This option allows the knowledgeable user to introduce a literal cut-off (instead of a relative, percentual threshold). If the option is set literal cut-off, will use **xFIToM** the value set in the threshold option as a literal cut-off, instead of as a relative threshold.

By default % of collection is set; i.e. threshold value will be used as relative threshold.

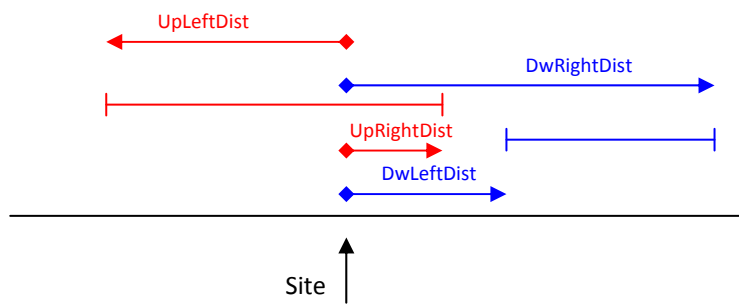
### 10 – Integrative factor

This option indicates whether **xFIToM** operates in normal (*no*) or integrative (*yes*) mode. By default *Integrative factor* is set to *No*, program will run in normal mode.

In integrative mode, **xFIToM** computes four mean values for sequence scores: local mean  $[-\max(X,Y), +\max(X,Y)]$ , upstream mean  $[-X, -x]$ , downstream mean  $[+y, +Y]$  and the global mean for all the sequence. The intervals can be freely specified by the user. For instance, one may compute the upstream mean from positions -200 to +50 of the site, and the downstream mean from positions +100 to +500 from site. The only restriction is that the intervals may not overlap. The local mean interval is always site-symmetrical and covers the maximum distance from the site specified by the user; in the above example it would cover the  $[-500, +500]$  interval.

After computing the pre-run mean (see above), **xFIToM** starts analyzing the genome with a look-ahead of  $\max(X,Y)$ . In this mode of operation, the cut-off value usually applied to putative sites is relaxed using a user-determined constant. After scanning the sequence, the sites with scores above the relaxed cutoff are re-evaluated using a correction factor that takes some of the computed means into account. Three different modes of score correction are possible: (upstream mean / global mean), (upstream mean / local mean),

(upstream mean / downstream mean). In all of them, the original score is multiplied by one of the above fractions.



11 – Further distance of current site to compute upstream mean (+/- X) -> [X,Y]

This parameter (UpLeftDistance) controls the farthest distance (from site under evaluation) in which the user wants to compute the upstream mean.

Default value of UpLeftDistance is -200, farthest distances from site to compute upstream mean.

12 – Closer distance of current site to compute upstream mean (+/- Y) -> [X,Y]

This parameter (UpRightDistance) controls the closest distance (from site under evaluation) in which the user wants to compute the upstream mean.

Default value of UpRightDistance is 50, closest distances from site to compute upstream mean.

13 – Further distance of current site to compute downstream mean (+/- I) -> [I,J]

This parameter (DwRightDistance) controls the farthest distance (from site under evaluation) in which the user wants to compute the downstream mean.

Default value of DwRightDistance is 850, farthest distances from site to compute downstream mean.

14 – Closer distance of current site to compute downstream mean (+/- J) -> [I,J]

This parameter (DwLeftDistance) controls the closest distance (from site under evaluation) in which the user wants to compute the downstream mean.

Default value of DwLeftDistance is 250, closest distances from site to compute downstream mean.

15 – Relaxation constant

This parameter controls the degree to which the normal threshold (specified either as relative or literal cutoff) is relaxed in integrative mode. Threshold relaxation is required for the integrative mode to introduce new information to ranking. Else, the method would only re-rank the sites already detected in normal operation. The default relaxation constant is 1.5. Bigger relaxation constants will provide the integrative mode with

additional freedom to choose sites based on their integrative component (a larger pool of candidate sites will be rescored), while smaller constants will reduce the contribution of the integrative factor. Care should be taken in specifying very big (>2) relaxation constants when analyzing large files, since this could result in a very large amount of sites selected for rescoring, which may lead to slow runs and, in extreme cases, out of memory errors.

#### *16 – Rescoring method*

The rescoring method option (1-3) allows the user to specify the method that will be applied to rescore candidate sites in integrative mode. Three methods are available in xFITOM to rescore sites, which differ on the fraction they apply as a multiplicative correction factor:

- 1 - Upstream mean / global mean
- 2 - Upstream mean / local mean
- 3 - Upstream mean / downstream mean

By combining this option with interval definition (options 11-14); many different re-evaluation strategies may be assayed. For instance, if one is interested in obtaining the local/global mean, a suitable upstream region is defined as the intended local mean, regardless of the downstream interval, and method 1 is used.

Default value of rescoring method is 2; i.e. scores will be reevaluated multiplying by the upstream/local mean ratio.

## References

- Berg, O. G. and P. H. von Hippel (1987). "Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters." J Mol Biol **193**(4): 723-50.
- Berg, O. G., R. B. Winter, et al. (1981). "Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory." Biochemistry **20**(24): 6929-48.
- Erill, I. and M. C. O'Neill (2009). "A reexamination of information theory-based methods for DNA-binding site identification." BMC Bioinformatics **10**(1): 57.
- Halford, S. E. and J. F. Marko (2004). "How do site-specific DNA-binding proteins find their targets?" Nucleic Acids Res **32**(10): 3040-3052.
- Hertz, G. Z., G. W. Hartzell, 3rd, et al. (1990). "Identification of consensus patterns in unaligned DNA sequences known to be functionally related." Comput Appl Biosci **6**(2): 81-92.
- O'Neill, M. C. (1989). "Consensus methods for finding and ranking DNA binding sites. Application to *Escherichia coli* promoters." J Mol Biol **207**(2): 301-10.
- O'Neill, M. C. (1998). "A general procedure for locating and analyzing protein-binding sequence motifs in nucleic acids." Proc Natl Acad Sci U S A **95**(18): 10710-5.
- O'Neill, M. C. (2003). "A procedure for identifying loosely conserved protein-binding DNA sequences." Methods Enzymol **370**: 229-37.
- Ricchetti, M., W. Metzger, et al. (1988). "One-Dimensional Diffusion of *Escherichia coli* DNA-Dependent RNA Polymerase: A Mechanism to Facilitate Promoter Location." Proc Natl Acad Sci U S A **85**(13): 4610-4614.
- Schneider, T. D. (1997). "Information Content of Individual Genetic Sequences." Journal of Theoretical Biology **189**(4): 427-441.
- Schneider, T. D., G. D. Stormo, et al. (1986). "Information content of binding sites on nucleotide sequences." J Mol Biol **188**(3): 415-31.
- Yellaboina, S., J. Seshadri, et al. (2004). "PredictRegulon: a web server for the prediction of the regulatory protein binding sites and operons in prokaryote genomes." Nucleic Acids Res **32**(suppl\_2): W318-320.