Swiss Institute of Bioinformatics

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

École Polytechnique Fédérale de Lausanne

Swiss Institute of Bioinformatics

TEMPLE v.1.0

Maria Litovchenko*[1] and Stefan Laurent†[2,3]

[1]Ludwig Maximilian University of Munich
[2]School of Life Sciences, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
[3]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

January 18, 2015

*maria.litovchenko@gmail.com
†stefan.laurent@epfl.ch

1 Citation and information

Citation and information TEMPLE: a bioinformatic tool to analyze population genetic diversity at transcription factor binding sites. Contact: `maria.litovchenko@gmail.com` and `stefan.laurent@epfl.ch`

2 Installation

2.1 Requirements

Temple can run on any operating system provided that java runtime environment (JRE) 1.7 or higher is installed.

2.2 Download

1. Download the compressed (rar) file at the following link: `http://jensenlab.epfl.ch/page-86730-en.html`
2. Uncompress the rar file and enter the folder
3. Double click on the file called TEMPLE.jar to launch the graphical version. (make sure to make TEMPLE.jar executable)

3 Graphical version of TEMPLE

3.1 Structure of input files in the graphical mode

1. The sequence file.

   The file containing the DNA alignment has to be uploaded with "File/Open sequence(s)". The alignment should be in the FastA format but the following information has to be added to the sequence names.

   Chromosome: (e.g. X, 2L, 19,...)

   Start coordinate: This can be the start coordinate of the sequence in an appropriate reference genome or simply set to "1" if a reference genome is not available.

   Stop coordinate: Coordinate of the last nucleotide of the sequence in the coordinate system of an appropriate reference sequence. Alternatively this can be set to the length of the sequence if an appropriate reference genome is not available.

   ID: ID of the sequence.

   These four information fields have to be specified in the following way: CHR_START_STOP_ID. Here are two examples of a valid sequence name:

   >X_1200000_1205000_Pop1seq35

   >chromosome_1_10000_PopSouthIndividual23.

   The sequences themselves can be written on one line or on multiple lines. If sequences are longer than 250kb please consider splitting them over

chr	start	stop	id	pwms
chrX	1	1000	region1	pho
scaffold34	150	950	geneX	all

Table 1: An example of region file.

multiple lines in order to speed up the execution time of the program. Note that TEMPLE does not recognizes the IUPAC ambiguity code. The only valid states in the FastA sequence are A,T,G,C,N,-, where "N" and "-" represents missing data and gaps, respectively. Please note that TEMPLE will not align the sequences. The user has to make sure that the DNA sequences in the sequence file represent a valid alignment. Allowed file extensions for the sequence file are .fas, .fasta., and, .afa.

Important note: names of the sequences must be unique

2. The region file.

   The "region file" can be uploaded from the menu by clicking on "File/Open region(s) file". This file contains the information on the genomic regions in which TFBS will be predicted and analyzed. Several regions can be defined on successive rows. The first line of the file has to be a header containing the following column names separated by tabulations or semicolons: chr, start, stop, id, pwms. All columns have to be present and specified. Here is a description of each column.

   Column 1 chr: The name of the chromosome

   Column 2 start: Start coordinate of the region in the alignment.

   Column 3 stop: Stop coordinate of the region in the alignment.

   Column 4 id: name of the region.

   Column 5 pwms: this column tells TEMPLE which position weight matrices (PWM) will be used to score the region(s). Possible values are: all or a list of PWMs names separated by commas (e.g. pho, zeste)

   Table 1 illustrates an example of a possible region file.

3. PWM file.

   The "PWM" file has to be uploaded from the menu by clicking on "File/Open PWM file". This file has to contain horizontal or vertical count position weight matrices. (The format is the same as the one provided by the "Fly Factor Survey" database (http://pgfe.umassmed.edu/ffs/). Similarly to the FastA format the first line has to start with the "greater than" sign (">") followed by the name of the PWM. The next four lines (or columns in the vertical format) contain the matrix itself. The PWM file can be used to upload multiple PWMs. Table 2 illustrates an example of a valid PWM.

3.2 Specifying sequence and region sets

Grouping sequences into sequence sets can be done from the menu by clicking on "Data/Specify sequence sets".

>pho															
A |	142	111	134	299	244	221	419	0	0	0	37	12	81	139	77
C |	75	159	204	26	59	52	0	0	0	0	383	44	62	208	146
G |	96	88	58	80	6	90	0	0	429	429	0	340	243	30	66
T |	84	64	33	24	120	66	10	429	0	0	9	33	43	49	115

Table 2: An example of PWM file.

1. Sequence sets

   Example:

   - Specify a name for the new set (see Figure 1)
   - Select sequences in the left panel using the mouse. Multiple sequences can be selected with the use of the "ctrl" and "shift" keys. You can also search for sequences which name share a common substring (for example "ZI" in figure 3). Just type in the searched string in the search box and press "search!" (see Figure 2).
   - Move all selected sequences to the right panel by clicking on the "»»>" button.
   - Save the sequence set by clicking on the "Submit" button.
   - Close the window or create a new sequence set by clicking on the "New" button. (A sequence set should not be created for the outgroup sequence)

2. Region sets

   Region sets can be defined from the menu by clicking on "Data/Specify region sets". and are created in the same way than sequence sets. Region sets can be used to restrict the analyses to a subset of the regions contained in the region file.

3.3 Saving and loading workspaces

Temple saves all settings relative to the sequence and region sets, the uploaded PWMs, as well as an index of the sequence file in workspaces. Saving and loading workspaces can be done with "File/Save workspace" and "File/Load workspace". For large alignments (>250kb), it is highly recommended to use of workspaces.

3.4 Single population analysis

A single population analysis can be launched from the menu by clicking on "Analysis/Single set". Five independent analyses using different sequence and region sets can be specified and conducted in parallel. Note that depending on the size of the uploaded datasets the time needed for the calculation to complete can vary from a few minutes to several hours.

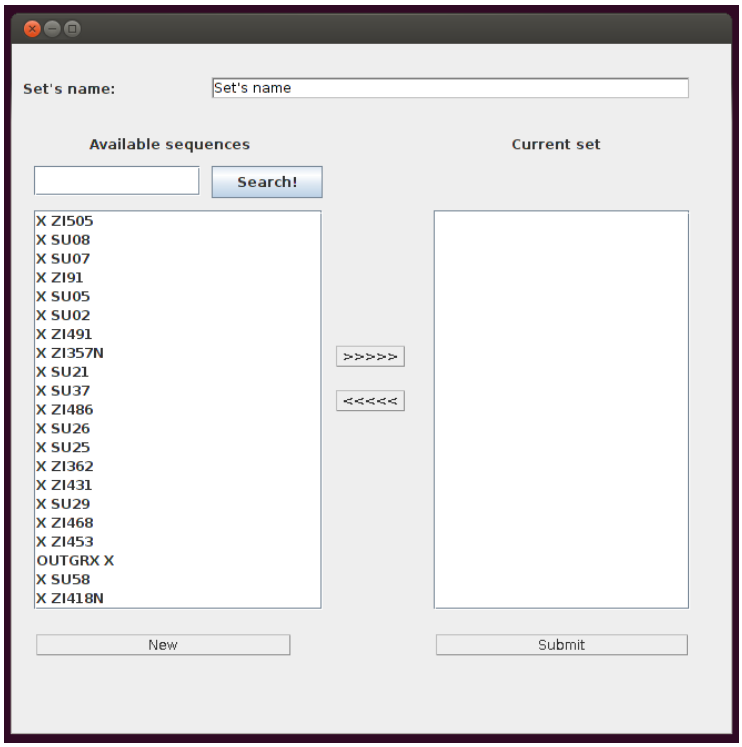

Figure 1: The sequence set definition window.

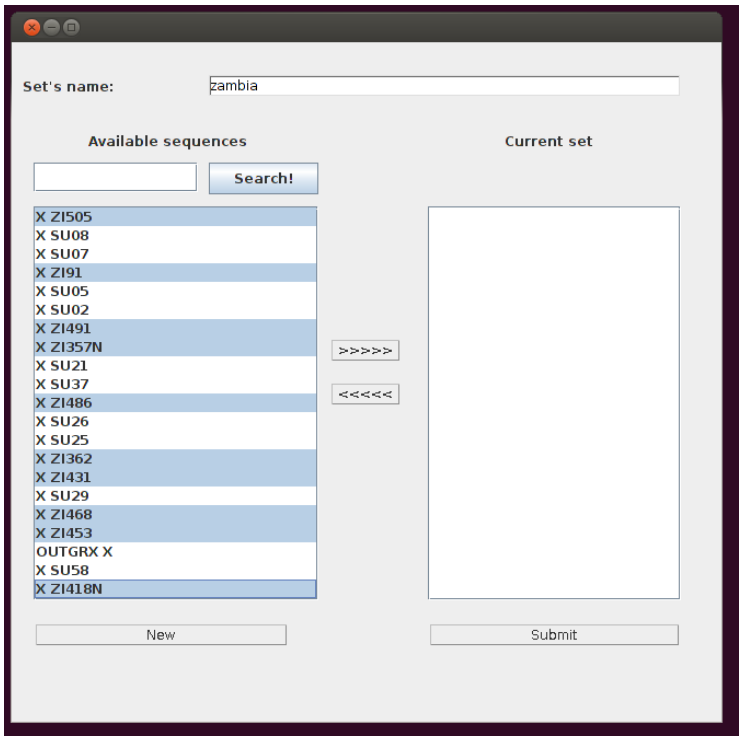

Figure 2: Creating a new sequence set.

3.5 Two population analysis

A two-population analysis can be launched from the menu by clicking on "Analysis/Double set". The settings are similar to the case of one population.

3.6 Visualization

After temple has successfully finished the calculations the program stores all results in the result folder (which path was specified by the user). Results can be visualized by clicking on "View/View results". This will open a window in which the user is prompted to specify the samples and the regions that he wishes to visualize. (In the case of a single population analysis the user is required to explicitly set the second sequence set to “empty”). If the analysis was conducted with an outgroup sequence the name of the outgroup can be defined at this stage. The user also has to choose which genomic region he wants to visualize. Finally, the user has to specify to the program the location of the .tfbs file (located in the result folder). Pressing on proceed should open the following window (this can take several minutes).The user has the possibility to highlight DNA regions for which a significant score was calculated by the TFBS prediction function. This can be done for up to five different PWMs using the different drop-down menus available on the upper part of the screen.

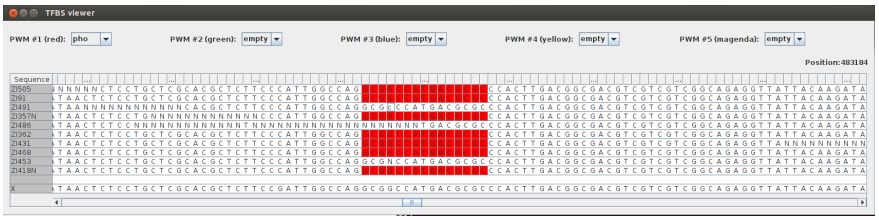

Figure 3: The visualization window

3.7 Diversity analysis

The main function of TEMPLE is to report and analyze the genetic diversity that occurs at TFBS loci. For TEMPLE, a TFBS loci is a loci at which at least one sequence has generated a significant PWM score. The genetic diversity identified at TFBS loci is reported in three different output files that represent three hierarchical levels at which diversity can be observed: the mutation, TFBS, and region level. In this three output files each row represents, a mutation, a tfbs, and a region, respectively. The three files have essentially the same format. The first lines provide a short summary of the dataset that was analyzed and are followed by the "results" section. Here is a description of this three files.

4 Command-line version of TEMPLE

TEMPLE can also be run in the command-line mode. The only differences with GUI mode described above are the arguments that are passed on the command line and the additional information that needs to be added to the alignments and region files.

4.1 Structure of input files

1. The sequence file.

   Overall, the input file containing sequences in case of one population analysis should be in commonly accepted fasta format. However, following changes should be made to the file:

   - The user should indicate to TEMPLE which sequences belong to which populations. This is done by grouping the sequences according to which population they belong to in the alignment file and by separating the two groups with a double slash symbol (//). Populations identifier can be specified after the "//". Listing 1 provides an example:

set	chr	start	stop	id	pwms
pop_north&pop_south	chrX	1	1000	region1	pho
pop_south	scaffold34	150	950	geneX	all

Table 3: An example of region file for command line version.

```
//pop_north
>chrX_12000_20000_northIndividual1
AGTGTGTGTA
>chrX_12000_20000_northIndividual2
AGTGTGTGTA
//pop_south
>chrX_12000_20000_southIndividual1
AGTGTGTGTA
>chrX_12000_20000_southIndividual2
AGTGTGTGTA
//
```

Listing 1: Example of sequence file for command line version.

- The sequence names should follow the same pattern as described in 3.1
- If an outgroup sequence is included, the name of the sequence should start with the letters "OUTGR". For the two population analysis, outgroup sequence should be imperatively located in the second group. Listing 2 provides an example:

```
//pop_north
>chrX_12000_20000_northIndividual1
AGTGTGTGTA
>chrX_12000_20000_northIndividual2
AGTGTGTGTA
//pop_south
>chrX_12000_20000_southIndividual1
AGTGTGTGTA
>chrX_12000_20000_southIndividual2
AGTGTGTGTA
>OUTGRchrX_12000_20000_sisterSpecies
AGTTTATGTA
//
```

Listing 2: Example of outgroup sequence set-up

Important note: file must end with /

2. The region file

   An additional column ("set") is needed in the region file when using the command-line version. In the one population analysis, this column should contain the id of the sequence set that has to be analyzed for a given region (e.g. pop_north in the example above). The ID should be the same as the one used in the alignment file. In the two population analysis the sequence set IDs have to be combined with the ampersand symbol ("&", Figure 3).

>pho	pval=0.001														
A |	142	111	134	299	244	221	419	0	0	0	37	12	81	139	77
C |	75	159	204	26	59	52	0	0	0	0	383	44	62	208	146
G |	96	88	58	80	6	90	0	0	429	429	0	340	243	30	66
T |	84	64	33	24	120	66	10	429	0	0	9	33	43	49	115
>pho1	score=100														
A |	142	111	134	299	244	221	419	0	0	0	37	12	81	139	77
C |	75	159	204	26	59	52	0	0	0	0	383	44	62	208	146
G |	96	88	58	80	6	90	0	0	429	429	0	340	243	30	66
T |	84	64	33	24	120	66	10	429	0	0	9	33	43	49	115

Table 4: An example of PWM file.

3. The PWM file

   The format of the PWM file is the same as in the GUI mode, but information about the significance threshold used for the motif prediction can be specified next to the line with the name of the sequence with use of key-words "pval" and "score", see table 4 for example. If no information is specified TEMPLE will assumes that the maximum ln(p-value) for printing scores equals the negative of the sample-size adjusted information content (-li option in patser).

4.2 Execution

Execution of command line version of the program is achieved by the following command:

java -jar gTemple.jar "`path/to/pwm/file`" "`path/to/fasta/file`" "`path/to/region/file`" "`number/of/threads`" "`threshold/for/missing/data`" "analysisMode"

where

- path/to/pwm/file is a location of the file containing the position weight matrices
- path/to/fasta/file is a location of the file containing the sequence alignments
- path/to/region/file is a location of the file containing the description of the regions to be analyzed
- number/of/threads is a number of java multi-threads requested for the analysis
- threshold/for/missing/data is a proportion of valid DNA bases (i.e. A,T,G,C) at a given polymorphic site below which the site is not analyzed
- analysisMode equals 1 for a one population analysis and 2 for a two-population analysis.

5 The output files

As a result, TEMPLE produces three output file: _MUTATION.csv, _TFBSunique.csv and _REGION.csv. Information, stored in the files, is described below.

5.1 The Mutation output file

1. Region

   The name of the region in which the mutation has been found

2. Chr

   The chromosome on which the mutation has been identified

3. Region_start

   The absolute position of the first nucleotide in the region

4. Region_stop

   The absolute position of the last nucleotide in the region.

5. PWM

   The name of the PWM as it is specified in the PWM file.

6. Position

   Position of the mutation in the coordinate system of the uploaded DNA sequence (as specified in the name of the sequence in the fasta file)

7. Pos_in_tfbs

   Position of the mutation within it's TFBS on the forward strand.

8. Reverse

   Indicates whether the TFBS has been predicted on the forward (TRUE) or reverse strand (FALSE).

9. Type

   Type of the genetic variant. Possible values are "snp" or "indel".

10. NumOfAlleles

    Number of alleles. The number of distinct allelic states observed at this position.

11. Length

    Length of the genetic variant. If it is a snp length is always one. For an indel it is the length of the indel.

12. Polarization

    Possible values are -1, 0, and 1."1" means that TEMPLE was able to use a valid outgroup sequence at this position to infer which of the two segregating alleles is the ancestral and which is the derived one."0" means that no valid outgroup sequence was available (i.e. missing data, or allelic state not present in ingroup) for this site in which case TEMPLE considers

that the derived allele is the allele with the lowest frequency. Finally, "-1" means that the site was not polarizable (e.g. more than 2 allelic states are present).

13. A1,A2,A3,A4

    The state of the observed alleles. If polarization is equal to 1 then A1 is the ancestral state (as inferred by the outgroup) and Al2 is the derived state. If polarization was not possible Al2 is the minor allele. If more than 2 alleles are present at a site, the order has no specific meaning.

14. Al1_pop1, Al2_pop1, Al3_pop1, Al4_pop1

    The absolute frequencies of the allelic states reported in Al1,Al,2,Al3, and Al4 in population 1 (population1 will have the name of the first sequence set that was defined). These column will be printed for the second population as well in a double population analysis.

15. NumbMisLines_pop1

    The number of sequences in the alignment with missing data at this position in the first population (A second column is present in a double population analysis).

16. Pi

    Nucleotide diversity calculated as 2pq, where p and q are the relative frequencies of the derived and ancestral allele, respectively.(Note that in a double analysis pi will be calculated on the pooled sequence sets)

17. Fst

    The Fst statistic as proposed by . Only for two populations-analyses.

18. EffectOnNucleotide

    The difference in TFBS scores caused by the derived allele at the nucleotide level. This is calculated as delta S = M(i,xderived) - M(i,xancestral), where M is the logodds pwm, i is the position of the SNP in the TFBS, and xderived and xancestral are the allelic states of the derived and ancestral alleles. If polarization is not possible the minor allele is considered to be the derived allele.

19. EffectOnTfbs

    This is the average effect of the derived allele on the TFBS scores. This statistic is useful when more than one mutation occur at a TFBS. It is calculated as the differences between the average scores of the TFBS carrying the derived and the ancestral alleles, respectively. Note that all non-significant scores, as well as TFBS with one or more missing sites (N) are set to 0 before calculation of the averages so that this statistic will always different from the "EffectOnNucleotide".

20. EffectOnRegion

    This is similar to the EffectOnTFBS column, but here the score of a single TFBS is replaced by the sum of all scores across the complete region.

Currently, only sequences with zero missing data are included in the analysis. This statistic can be used to identify mutations that contribute to an increase or decrease of binding affinity over the whole region (as defined in the region file) , taking into account potential compensatory mutations.

5.2 The TFBS output

Columns that have the same name as in the mutation output file contain the same information and are not re-described here.

1. Threshold

   This is the cutoff value used by the PWM scoring method to identify significant TFBS. It is calculated following . Temple assumes that the maximum ln(p-value) for printing scores equals the negative of the sample-size adjusted information content.

2. Len

   The length of the predicted TFBS.

3. NumMut

   The number of polymorphic sites at this TFBS loci

4. SumPi

   Nucleotide diversity calculated as the sum of 2pq across all mutations within the TFBS.

5. mean_S_pop1

   The average of the distribution of TFBS score in population 1. Non-significant scores and scores of TFBS containing missing data are set to 0 in the distribution of scores.

6. sd_S_pop1

   The standard deviation of the distribution of TFBS score in population 1

7. S_Qst

   The equivalent of the Fst statistic for a continuous phenotype. Here we consider the PWM score at the TFBS level as a truncated continuous phenotype for which values below the PWM threshold are set to zero. The S_Qst correspond to proportion of the total variance in truncated scores that is explained by the population effect. (The truncation is there to avoid that variation among non-significant scores affects the Qst statistic). Scores of TFBS affected by missing data are set to 0 as well.

5.3 The region output

1. SumPiTFBS

   The nucleotide diversity measured at TFBS site in this region. Calculated as the sum of 2pq across all polymorphic sites in all TFBS loci in this region.

2. LenTFBS

   The sum of the length of all TFBS in this region. This can be used to measure the per site nucleotide diversity.

3. NumTFBS

   The number of TFBS loci that have been predicted in this region

4. NumPolyTFBS

   The number of polymorphic TFBS_loci predicted in this region

5. mean_RS_pop1

   The mean of the distribution of region scores in sequence set 1. Region scores are calculated by summing over all significant TFBS scores in a region for a given sequence. Sequences for which one or more TFBS contain missing data ("N") will not be taken into account in the distribution of scores. (This is done to avoid that missing data affects the variation in score).

6. sd_RS_pop1

   The standard deviation of the distribution of region scores (RS) in sequence set 1.

7. RS_Qst

   This statistic is similar to the S_Qst value but is applied to region scores instead of TFBS score. This statistic can be used to identify regions that have large differences in score between populations.

6 Supplementary: Algorithm description

Bioinformatics predictions of TFBS rely on the assumption that all sequences that bind to a given DNA-binding protein share a certain level of similarity. Investigating the degree of similarity within a set of functionally related sequences starts by collecting DNA sequences that have been empirically shown to bind a given DNA-binding protein (e.g. genome-wide profiling experiments). The second step is to create an alignment of these sequences that maximizes sequence conservation. This alignment is then used to identify a sub-region characterized by a statistically significant high level of conservation. This region is then defined as the motif to which the DNA-binding protein binds. Although motifs consist in more or less conserved DNA sequences they usually contain some variation, such that these motifs cannot be fully represented by a single consensus sequence. A more appropriate way of representing such a motif is an alignment matrix (or count matrix). An example of such a matrix is given in Fig.1. This matrix contains the number of times, nxi, that a letter x is observed at position i in the alignment. In this matrix, for example, the positions seven to ten are highly conserved. This is the type of matrices that has to be provided as input to the program. When searching for occurrences of TFBS in a genome or in a sub-genomic region, these matrices can be used to identify substrings that are similar to the motif described in the matrix. In practice, this is done by transforming the alignment matrix into a weight matrix, whose elements are

the weights used to construct a similarity score to the motifs for every position of the investigated DNA sequence . Scores that are higher than a previously defined statistical threshold are reported and considered as bioinformatically predicted TFBS. In the following section we describe how the program transforms alignment matrices into weight matrices, how weight matrices are used to calculate a similarity score, how it calculates the probability distributions of score for every weight matrix, and which possible options are provided to the user to define a statistical threshold for TFBS predictions.

6.1 Transformation of alignment matrices into weight matrices

Let's consider a position weight matrix M indexed by (1...m) x 4, where m is the length of the motif. The coefficients M(i,x) give the scores at position i in [1,m] for the letter x in A,T,G,C. Mark follows [1] and calculates M(i,x) as follows:

$$M(i,x) \quad = \quad \ln \frac{(n_{xi} + p_x) \quad / \quad (N \quad + \quad 1)}{p_x} \approx ln \frac{f_{xi}}{p_x} \tag{1}$$

where n_{xi} is the number of times a letter x is observed in the alignment matrix, p_x is the apriori probability of the letter x, and N is the total number of sequences that was used to create the alignment matrix. The users should refer to Herz and Stormo (1999) for more detailed information. TEMPLE calculates p_x from the local base-pair composition of the currently analyzed genomic region.

6.2 Scoring function

Scoring DNA sequences given a position weight matrix is done using a sliding-window approach. We start at the first base pair of the sequence and calculate

$$Score(u, \quad M) \quad = \quad \sum_{i \quad = \quad 1}^{m} M(i, \quad u_i) \tag{2}$$

Where u is a substring of size m, M is a position weight matrix, and u_i denotes the nucleotide found at position i in u. Once the score is calculated we calculate the next scores at each successive position in the sequence smaller or equal to L − m + 1, where L is the length of the analyzed region. Note that when u_i = "N" (i.e. missing information), M(i, u_i) = 0. Scores associated with a significant p-value are reported in the output files. The reverse strand is scored by scoring the forward strand with a reversed-complemented PWM.

6.3 Calculation of p-values and determination of the cutoff score

To asses the significance of a given score, it is necessary to calculate the probability that a score equal or higher occurs under a background model characterized by a randomized DNA sequence with the same a priori probabilities for the different nucleotidic states (p_x). To do this, we followed Staden (1989) and used probability-generating functions to calculate the complete probability distribution across all possible scores given a position weight matrix (M) and p_x. We

recommend the users to read Staden (1989) as well as for a complete description of the algorithm.

Significance thresholds for motif predictions can be defined from the menu in "Data/Specify thresholds/pvalue for PWM". Here the user has access to three alternatives: set a threshold on the p-value, set a threshold on the score value itself, or let TEMPLE calculate a threshold automatically (default behavior, "Auto"). When the automatic threshold calculation is chosen, TEMPLE assumes that the maximum ln(p-value) for printing scores equals the negative of the sample-size adjusted information content (-li option in patser).

References

[1] Gerald Z Hertz and Gary D. Stormo. Identifying dna and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15(7):563–577, 1999.

Contents