Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W253–W258***
doi:10.1093/nar/gkm272

STAMP: a web tool for exploring DNA-binding motif similarities

Shaun Mahony[1] and Panayiotis V. Benos[1,2,*]

[1]Department of Computational Biology, School of Medicine, University of Pittsburgh and [2]Department of Human Genetics, Graduate School of Public Health, and University of Pittsburgh Cancer Institute, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA

Received January 31, 2007; Accepted April 10, 2007

ABSTRACT

STAMP is a newly developed web server that is designed to support the study of DNA-binding motifs. STAMP may be used to query motifs against databases of known motifs; the software aligns input motifs against the chosen database (or alternatively against a user-provided dataset), and lists of the highest-scoring matches are returned. Such similarity-search functionality is expected to facilitate the identification of transcription factors that potentially interact with newly discovered motifs. STAMP also automatically builds multiple alignments, familial binding profiles and similarity trees when more than one motif is inputted. These functions are expected to enable evolutionary studies on sets of related motifs and fixed-order regulatory modules, as well as illustrating similarities and redundancies within the input motif collection. STAMP is a highly flexible alignment platform, allowing users to 'mix-and-match' between various implemented comparison metrics, alignment methods (local or global, gapped or ungapped), multiple alignment strategies and tree-building methods. Motifs may be inputted as frequency matrices (in many of the commonly used formats), consensus sequences, or alignments of known binding sites. STAMP also directly accepts the output files from 12 supported motif-finders, enabling quick interpretation of motif-discovery analyses. STAMP is available at http://www.benoslab.pitt.edu/stamp

INTRODUCTION

'Position-specific scoring matrices' (PSSMs) and their derivatives have become the standard representation of a transcription factor's (TF) DNA-binding preference. For example, experimentally derived DNA-binding preferences for a growing number of TFs are stored as frequency matrices in databases such as JASPAR (1) and TRANSFAC (2). In addition, most *de novo* motif-finding software tools report statistically over-represented degenerate sequence features in the form of frequency matrices or consensus sequences.

Motif-discovery is often one of the first steps performed during computational analysis of gene-regulation. For instance, researchers often wish to discover over-represented motifs that are common to sets of genes with similar expression patterns. However, interpretation of the output from motif-finders is often daunting; many distinct motifs may be reported with little or no indication as to whether each may potentially possesses regulatory function. Furthermore, no information is provided about the TF protein that may bind to them. It is therefore surprising that few tools currently exist that can assess similarity between novel, computationally identified motifs and the known motifs stored in the databases. Available tools [such as T-Reg Comparator (3) and MACO (4)] currently allow for only a single type of alignment method, which may not be suitable for all database searches, and none support the direct analysis of motif-finder output files.

Recently, a number of studies have focused on the evolution of binding preference amongst related TFs. For example, generalized models of the binding preferences from a group of structurally related TFs have been described (5). Such 'familial binding profiles' (FBPs) have been shown to have wide applicability in improving the performance of motif-finders (5,6) and in predicting the structural class of the TF associated with novel motifs (5,7). Other studies have shown evolutionary conservation and change in fixed-order *cis*-regulatory modules (e.g. in the SXY modules controlling vertebrate MHC gene expression (8)). Currently, however, there is no publicly available software to support evolutionary analyses of DNA-binding motifs and facilitate the study of FBPs.

In response to the gap in the current bio-informatics software repertoire outlined above, the

*To whom correspondence should be addressed. Tel: +412 648 3315; Fax: +412 648 3163; Email: benos@pitt.edu or shaun.mahony@ccbb.pitt.edu

STAMP web server aims to provide a platform for 'BLAST-like' database searching and 'ClustalW-like' multiple alignment and tree building for DNA-binding frequency matrices and motifs. Instead of limiting analyses to a single ungapped alignment strategy, STAMP allows various combinations between the implemented scoring metrics, pairwise alignment methods, gap penalties, multiple alignment strategies and tree-building algorithms. The web server accepts many commonly used motif and frequency matrix formats, and in addition allows the uploading of entire output files from 12 supported motif-finders. STAMP therefore offers a highly flexible and comprehensive toolbox for the study of relationships between TF-binding motifs.

MATERIALS AND METHODS

Pairwise comparison and alignment of motifs

At the core of STAMP's functionality is the efficient comparison and alignment of two motifs. Two motifs can be aligned using Needleman–Wunsch (global) or Smith–Waterman (local) alignment methods, based on column comparison scores calculated by one of the five supported distance metrics: (i) Pearson's correlation coefficient (9), (ii) Kullback–Leibler information content (3), **b** sum of squared distances (5), (iv) average log-likelihood ratio (ALLR) (10) and (v) ALLR with a lower limit of –2 imposed on the score. The latter option is provided as an attempt to ease the negative effect the ALLR scoring function has on the motif alignment due to its highly skewed scores (7).

A special ungapped type of Smith–Waterman local alignment is also provided, where the motif 'cores' (defined as the four most informative adjacent matrix columns) are aligned before extending the alignment. Various gap-opening penalty options are also offered, and differ according to the column-comparison metric employed. The gap extension penalty is currently set to half of the value of the gap-opening penalty. For Smith–Waterman alignments, users may choose to only recover alignments that overlap by at least three matrix positions, and can require that the local alignment be extrapolated to the matrix edges.

In order to avoid length biases when comparing motifs of different lengths, we used the method of Sandelin and Wasserman for the calculation of empirical *P*-values based on simulated PSSM models (5). Construction of a dataset of 10 000 simulated PSSMs that reflect the properties of PSSMs in the JASPAR database was performed as described by Sandelin and Wasserman (http://fork head2.cgb.ki.se/jaspar/additional).

Multiple motif alignment and phylogenetic tree construction

Users may choose between two motif multiple alignment strategies: 'progressive profile alignment' and 'iterative refinement'. In progressive profile alignment, the multiple alignment is built up by progressively aligning the nodes on an approximate guide tree in order of decreasing similarity. Iterative refinement initializes the alignment using the most similar pair of input motifs, and progressively adds the remaining motifs according to similarity to a profile based on the current alignment. Once the initial alignment is built, each motif is removed from the alignment in turn and is realigned to a profile based on the other aligned motifs. Gaps are encouraged to open in the same positions in the multiple alignments by negatively weighting the gap penalties in positions of the multiple alignment that already contain gaps.

Finally, users may choose between two tree-building algorithms; an agglomerative method [UPGMA (11)] and a divisive method that is based on a self-organizing tree algorithm [SOTA (12)]. UPGMA begins by assigning each input motif its own leaf node. At each time-step, the two nodes with the maximum average pairwise similarity are joined. The tree is built up through successive combinations of nodes until only one node (the root) remains. SOTA follows the opposite strategy. The tree is initialized with only one node (the root), which contains a rough alignment of all input PSSMs, and the node model is generated from this alignment. The root node then produces two identical offspring leaf nodes. During each time-step, the algorithm assigns the PSSMs to their most similar leaf nodes and then allows the node model to be updated in accordance with their current contents. SOTA also allows for small contributions from neighboring nodes during the update step. These contributions are designed to keep neighboring nodes similar. After a number of time-steps, the node with the highest degree of dissimilarity amongst its members is allowed to produce two identical offspring nodes. This competitive learning scheme continues until each leaf node contains a single PSSM.

Database matching

Besides motif alignments and tree-building construction, STAMP automatically queries each of the input motifs against a user-specified motif database to identify their 'best matching' known motifs. The following motif databases are currently supported: (i) JASPAR, (ii) TRANSFAC, (iii) *Saccharomyces cerevisiae* 'regulatory code' motifs [predicted by Harbison *et al.* (13) and MacIsaac *et al.* (14)], (iv) *Drosophila* motifs [DNase I footprinting data from (15), motifs generated by Dan Pollard], (v) DPInteract *Escherichia coli* motifs (16) and (vi) RegTransBase prokaryotic motifs (17). Alternatively, users may upload their own dataset of motifs to query the input motifs against. Users may choose to get listings of 1, 5 or 10 of the best-matching motifs in the queried database.

Input data types and formats

STAMP accepts queries of one or more motifs (no maximum query size is currently enforced). In order to enhance accessibility, the web server supports a wide variety of motif input formats. For example, users may input DNA-binding motifs as collections of position frequency matrices (also known as count matrices or PSSMs) in TRANSFAC, JASPAR, MEME, or 'Raw count' formats. Motifs may also be entered as consensus sequences in the IUPAC degenerate sequence alphabet,

or as multiple alignments of sample-binding sites. Users do not have to tell the system which data format is being used in a query; the system automatically senses if a supported format has been entered. Users may also mix-and-match input formats in a single run of the platform.

Alternatively, users may upload the entire output files from a number of supported *de novo* motif-finders. This option is expected to be useful to those users who wish to interpret the results of a DNA motif-finding analysis. STAMP can be used to match the motif-finder output against databases of known motifs, and can also illustrate similarities between the discovered motifs [some motif-finders, such as BioProspector (18), may report multiple copies of similar motifs in any given run]. Currently supported STAMP input formats include the output files from motif-finder programs like SOMBRERO (19), BioProspector (18), MDScan (20), AlignACE (21), MEME (22), Weeder (23), MotifSampler (24), YMF (25), ANN-Spec (26), Consensus (27), Improbizer and Co-Bind (28). Examples of all supported formats are illustrated on STAMP's help web page.

Parameters

The users may specify any combination of the alignment parameters (comparison metric, gap penalty, alignment and tree-building strategy) and search database described above. Users may also specify the 'information content' for edge trimming. Motifs predicted by many motif-finders or stored in the databases often contain a 'core' region of high information content flanked by low information-content columns at the edges. Many researchers assume that most of these flanking columns are irrelevant to the protein–DNA interaction. Whether or not this assumption is true, STAMP allows the option of trimming these low information-content edges from the input motifs. Since STAMP's motif alignment *P*-value calculation is dependent on the length of the compared motifs, removing the low information-content edges can assist accurate alignment. STAMP allows the user to choose an information content threshold (between 0 and 1) for the purposes of excluding edge columns. The motif will not be shortened below the minimum motif length of four columns.

Implementation

The STAMP platform is written in C++, and is modularly designed to allow any combination of the implemented column comparison metrics, alignment methods, multiple alignment strategies and tree-building algorithms. Some of the implemented algorithms make use of functions provided by the open-source GNU scientific library (http://www.gnu.org/software/gsl). The web server is written in PHP with supporting scripts written in Perl. Sequence logos are generated using 'WebLogo' (29), and the similarity tree is drawn using the Phylip software package (30). Conversion of the results pages to Portable Document Format (PDF) is achieved using the open-source htmldoc program (http://www.htmldoc.org/). STAMP does not have excessive computing requirements. In its current deployment (on a Dell PowerEdge 2650 server with 2.8 GHz dual Xeon processors), and under typical server loading conditions, STAMP processes a typical input dataset (e.g. 10 motifs) in ~5 s.

RESULTS AND DISCUSSION

STAMP functionality

STAMP's typical functionality is demonstrated in Figure 1 using a selection of four bHLH structural class motifs taken from the JASPAR database. The tested motifs are NHLH1 (MA0048), TAL1-TCF3 (MA0091), MAX (MA0058) and USF1 (MA0093). Note that the matrices are of different lengths. The four JASPAR matrices are pasted into the motif input box, and we choose to perform an ungapped Smith–Waterman alignment using Pearson's correlation coefficient, an iterative refinement multiple alignment, and a UPGMA tree. In this example, we choose to match the input motifs against the MacIsaac *et al.* (14) dataset of *S. cerevisiae* motifs. STAMP's results page is displayed as a HTML document, but may also be downloaded as a PDF. As illustrated in Figure 1, STAMP results include the multiple alignment of the input motifs (when two or more motifs are provided in the input), a similarity tree (when three or more motifs are provided in the input), and a ranked list of matches in the chosen dataset for each input motif.

Although the multiple alignment algorithm is carried out on the original inputted frequency matrices, the resulting motif multiple alignment is displayed in IUPAC consensus sequence format (Figure 1A). The multiple alignment is accompanied by the generalized profile, which represents the average profile of the multiple alignment. Generalized profiles (including FBPs) are useful when studying the binding properties common to a set of related motifs. For example, the familial binding motif in Figure 1A illustrates the 'CA'- and 'TG'- binding positions that are shared between all four input motifs. Clicking on the 'Alignment Profile' hyperlink allows the generalized profile's frequency matrix to be downloaded.

A simple figure representing the motif tree is also displayed on the results page (Figure 1B). In the example in Figure 1B, the UPGMA tree successfully separates the bHLH motifs (NHLH1 and TAL-TCF3) from the bHLH-ZIP motifs (MAX and USF1), since the binding preference of these two subclasses are distinct ('CAGCTG' and 'CACGTG', respectively). The logos of the input motifs are plotted beside the tree figure.

Finally, the results of the database queries are shown (Figure 1C for an example for the USF1 motif). For each input motif, STAMP lists a user-defined number of best-matching motifs from the user-specified database. The results reproduce the name and sequence logo of the database hit, along with the pairwise alignment (in consensus sequence format) and the E-value corresponding to the alignment. From the partial results reproduced in Figure 1C, it may be seen that the human

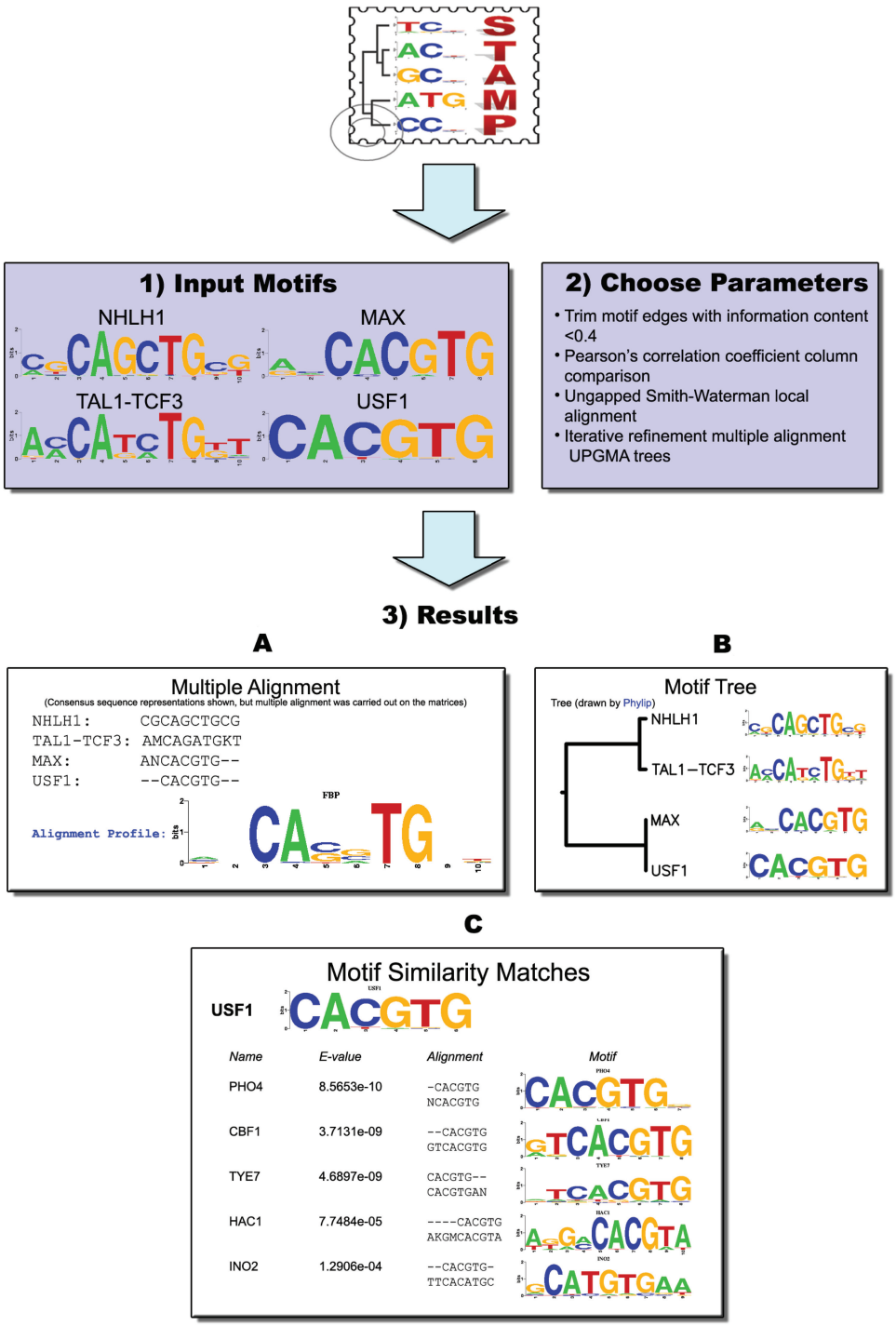

Figure 1. An illustration of STAMP's analysis for four bHLH DNA-binding motifs. Motifs are inputted as PSSM models or consensus sequences. For illustration purposes, this figure represents them as LOGOs. The user-specified parameters include the distance metric to be used, and the alignment and tree-building strategy of choice. In the output STAMP reports the multiple alignment of the input motifs (**A**), the corresponding distance tree (**B**) and the best similarity matches against the database of choice (**C**).

USF1 motif has a number of close matches in the dataset of known *S. cerevisiae* motifs.

Other uses for STAMP

Aside from STAMP's more obvious database searching and multiple alignment functionality (Figure 1), the web server may also be used to support various other types of analyses. Some examples follow.

Obtain information about the identity of a TF that binds to a DNA motif. When *de novo* pattern discovery methods are used to analyze sets of unaligned DNA sequences (typically, the promoters of co-expressed genes identified

through some high-throughput gene expression technique), they usually identify a number of statistically significant DNA motifs. In general, the TF that binds to these motifs is unknown. With the database search functionality of STAMP, the user can now find for each new motif its 'closest known relative' motif or its most likely FBP membership.

Furthermore, it has been shown that TFs from the same structural class often share similar binding preferences (5). This fact has been used to allow prediction of the structural class of TFs associated with novel DNA motifs, either by comparing the novel motif to generalized FBPs (5) or by using the best hit in a well-represented database to provide the prediction (7). Thus, even if the TF that recognizes a newly discovered motif is currently unknown, using STAMP to compare novel motifs against the JASPAR, TRANSFAC or the supported sets of FBPs can provide information about the structure of the associated TF.

FBP construction. FBPs are generalized models of DNA binding for TFs that share structural similarities (5). Their uses include the identification of the structural group of the TF that recognizes a newly discovered DNA motif and utilization as priors for DNA pattern discovery algorithms (5). FBPs are constructed through multiple alignments of structurally related DNA-binding motifs. STAMP automatically generates a generalized profile from the final multiple alignment of the input motifs. An FBP for particular TF structural class or sub-class may therefore be constructed simply by providing STAMP with a collection of the corresponding motifs. The web server allows the exploration of different distance metrics and alignment strategies in order to construct optimal FBPs. In addition, STAMP's construction of hierarchical motif trees can be used to guide the definition of structural class subfamilies if more specific FBPs are required.

Future improvements

We are interested in expanding the number of supported databases of known motifs that STAMP can query, and we will try to respond to any specific suggestions in respect to this. Similarly, the number of supported motif input formats may be expanded in the future. We would encourage the developers of any currently over-looked motif-finders to provide us with sample output files, making us aware of any unique properties of the output format that distinguishes it from other motif-finder output. We also hope to incorporate other column comparison metrics, alignment methods and tree-building algorithms into the STAMP platform in the future.

CONCLUSIONS

STAMP is the first web server that facilitates multiple alignment and tree-building for collections of DNA-binding motifs. STAMP therefore aims to provide a platform for the evolutionary study of TF-binding motifs, just as ClustalW (31) and similar tools provide a platform for the evolutionary analysis of sequence information. A small number of other programs currently provide some of STAMP's database search functionality, including T-Reg Comparator (3), MACO (4) and the matrix query component in JASPAR (1). However, limited numbers of input formats are supported by these services, and each supports only a single-alignment strategy. In contrast, STAMP allows the user to mix-and-match between the five supported column comparison metrics, the three supported pairwise alignment methods and various gap penalties. Many different input formats are supported, and the user may directly upload the entire output files from 12 supported *de novo* motif-finders. In addition, a growing collection of general and species-specific known motif databases may be queried using STAMP.

We are confident that STAMP will be a useful resource for future studies of motif evolution, as well as allowing greater power in interpreting the often voluminous result files generated by *de novo* motif-finders.

ACKNOWLEDGEMENTS

This work was supported by NIH grants RR014214 and NO1 AI-50018 and NSF grant MCB0316255. P.V.B. was also supported by NIH grant 1R01LM007994-01 and TATRC/DoD USAMRAA Prime Award W81XWH-05-2-0066. Funding to pay the Open Access publication charges for this article was provided by NSF (grant no.: MCB0316255).

Conflict of interest statement. None declared.

REFERENCES

1. Sandelin,A., Alkema,W., Engstrom,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.
2. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
3. Roepcke,S., Grossmann,S., Rahmann,S. and Vingron,M. (2005) T-Reg Comparator: an analysis tool for the comparison of position weight matrices. *Nucleic Acids Res.*, **33**, W438–W441.
4. Su,G., Mao,B. and Wang,J. (2006) MACO: a gapped-alignment scoring tool for comparing transcription factor binding sites. *In Silico Biol.*, **6**, 307–310.
5. Sandelin,A. and Wasserman,W.W. (2004) Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J. Mol. Biol.*, **338**, 207–215.
6. Macisaac,K.D., Gordon,D.B., Nekludova,L., Odom,D.T., Schreiber,J., Gifford,D.K., Young,R.A. and Fraenkel,E. (2006) A hypothesis-based approach for identifying the binding specificity of regulatory proteins from chromatin immunoprecipitation data. *Bioinformatics*, **22**, 423–429.
7. Mahony,S., Auron,P. and Benos,P.V. (2007) DNA Familial Binding Profiles Made Easy: Comparison of Various Motif Alignment and Clustering Strategies. *PLoS Comput Biol,* **3**, *e61*.
8. Belov,K., Deakin,J.E., Papenfuss,A.T., Baker,M.L., Melman,S.D., Siddle,H.V., Gouin,N., Goode,D.L., Sargeant,T.J. *et al.* (2006) Reconstructing an ancestral mammalian immune supercomplex from a marsupial major histocompatibility complex. *PLoS Biol.*, **4**, e46.

9. Pietrokovski,S. (1996) Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res.*, **24**, 3836–3845.
10. Wang,T. and Stormo,G.D. (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, **19**, 2369–2380.
11. Sokal,R.R. and Michener,C.D. (1958) A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.*, **28**, 1409–1438.
12. Dopazo,J. and Carazo,J.M. (1997) Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree. *J. Mol. Evol.*, **44**, 226–233.
13. Harbison,C.T., Gordon,D.B., Lee,T.I., Rinaldi,N.J., Macisaac,K.D., Danford,T.W., Hannett,N.M., Tagne,J.B., Reynolds,D.B. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.
14. MacIsaac,K.D., Wang,T., Gordon,D.B., Gifford,D.K., Stormo,G.D. and Fraenkel,E. (2006) An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics*, **7**, 113.
15. Bergman,C.M., Carlson,J.W. and Celniker,S.E. (2005) Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster. *Bioinformatics*, **21**, 1747–1749.
16. Robison,K., McGuire,A.M. and Church,G.M. (1998) A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome. *J. Mol. Biol.*, **284**, 241–254.
17. 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.
18. Liu,X., Brutlag,D.L. and Liu,J.S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.,* 127–138.
19. Mahony,S., Golden,A., Smith,T.J. and Benos,P.V. (2005) Improved detection of DNA motifs using a self-organized clustering of familial binding motifs. *Bioinformatics*, **21**, i283–i291.
20. Liu,X.S., Brutlag,D.L. and Liu,J.S. (2002) An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.*, **20**, 835–839.
21. Hughes,J.D., Estep,P.W., Tavazoie,S. and Church,G.M. (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J. Mol. Biol.*, **296**, 1205–1214.
22. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Pro. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
23. Pavesi,G., Mereghetti,P., Mauri,G. and Pesole,G. (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.*, **32**, W199–W203.
24. Thijs,G., Lescot,M., Marchal,K., Rombauts,S., De Moor,B., Rouze,P. and Moreau,Y. (2001) A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics*, **17**, 1113–1122.
25. Sinha,S. and Tompa,M. (2003) YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.*, **31**, 3586–3588.
26. Workman,C.T. and Stormo,G.D. (2000) ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac. Symp. Biocomput.*467–478.
27. Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.
28. GuhaThakurta,D. and Stormo,G.D. (2001) Identifying target sites for cooperatively binding factors. *Bioinformatics*, **17**, 608–621.
29. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
30. Felsenstein,J. (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*, **5**, 164–166.
31. Chenna,R., Sugawara,H., Koike,T., Lopez,R., Gibson,T.J., Higgins,D.G. and Thompson,J.D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.*, **31**, 3497–3500.