Published online 27 November 2006

Nucleic Acids Research, 2007, Vol. 35, Database issue **D127–D131**
doi:10.1093/nar/gkl857

SwissRegulon: a database of genome-wide annotations of regulatory sites

Mikhail Pachkov[1,2], Ionas Erb[1,2], Nacho Molina[1,2] and Erik van Nimwegen[1,2,*]

[1]Biozentrum, The University of Basel, Klingelbergstrasse 50/70, 4056-CH, Basel, Switzerland and [2]Swiss Institute of Bioinformatics, Switzerland

Received August 14, 2006; Revised September 17, 2006; Accepted October 1, 2006

ABSTRACT

SwissRegulon (http://www.swissregulon.unibas.ch) is a database containing genome-wide annotations of regulatory sites in the intergenic regions of genomes. The regulatory site annotations are produced using a number of recently developed algorithms that operate on multiple alignments of orthologous intergenic regions from related genomes in combination with, whenever available, known sites from the literature, and ChIP-on-chip binding data. Currently SwissRegulon contains annotations for yeast and 17 prokaryotic genomes. The database provides information about the sequence, location, orientation, posterior probability and, whenever available, binding factor of each annotated site. To enable easy viewing of the regulatory site annotations in the context of other features annotated on the genomes, the sites are displayed using the GBrowse genome browser interface and can be queried based on any annotated genomic feature. The database can also be queried for regulons, i.e. sites bound by a common factor.

INTRODUCTION

Regulation of the rate of transcription initiation is one of the main mechanisms through which cells regulate the expression of proteins encoded in their genomes. Transcription regulation is generally implemented through the sequence-specific binding of transcription factors (TFs) to target sites in the DNA, which are most often in the intergenic region upstream of the regulated gene.

The sequence segments recognized by TFs are generally short, i.e. typically ∼20 bp for prokaryotic TFs and ∼10 bp for eukaryotic TFs, and are normally degenerate. In spite of decades of extensive experimental work the number of experimentally known binding sites accounts only for a small fraction of the total number of functional sites that likely exist. For example, probably the most extensive data are available for *Escherichia coli* with ∼1000 sites that have experimental support (1), and for *Sacchromyces cerevisiae* with a few hundred sites that have direct experimental support (2). However, even for *E.coli* this constitutes probably less than one-fifth of all binding sites that exist genome wide, and only about a third of the ∼300 TFs in *E.coli* are represented with at least one binding site.

Computational approaches for inferring transcription factor binding sites stretch back almost two decades (3–5). However, only with the recent advent of large numbers of fully sequenced genomes, and the availability of genome-wide gene expression and chromatin immuno precipitation data has it become computationally feasible to comprehensively annotate regulatory sites genome-wide. For example, several approaches have been developed recently that identify regulatory sites by searching for significantly conserved sequence segments within multiple alignments of orthologous intergenic regions of related genomes (6–11). In this context several yeast species were sequenced recently (12,13) with the aim of identifying regulatory sites genome-wide. In addition to comparative genomic approaches large-scale ChIP-on-chip experiments have been undertaken recently in yeast to determine the intergenic regions bound by over 100 TFs (14,15). Computational approaches that combine comparative genomic analysis of orthologous intergenic regions with the analysis of these large-scale ChIP-on-chip data have led to the first comprehensive genome-wide annotations of binding sites in yeast (11,15–17).

REGULATORY SITE ANNOTATION METHODS

The methods that we use to produce the regulatory-site annotation for a given genome depend on the amounts and kinds of data that are available for that organism. For most organisms currently in SwissRegulon the only available data consist of the sequence of the genome and the genome sequences of related organisms. For these organisms our annotations are based on a careful comparison of orthologous intergenic regions from sets of related organisms as described below. For some genomes there are collections of known binding sites and we use these to build position-specific weight matrices (WMs) that represent the sequence-specificities of

*To whom correspondence should be addressed. Tel: +41 61 267 1576; Fax: +41 61 267 1584; Email: erik.vannimwegen@unibas.ch the TFs for which sites are available. For yeast there are also comprehensive ChIP-on-chip binding data available and we use these in combination with known sites to build a large set of WM models of yeast TFs. We use these sets of WMs to scan multiple alignments of orthologous intergenic regions genome-wide using the algorithm MotEvo (16).

IRUS: Intergenic Regions Under Selection

At the time of writing there are 354 complete microbial genomes that are available from the NCBI database (18). For all but a handful of these genomes there are no known regulatory sites, nor any ChIP-on-chip data available. However, for almost any genome in this collection one can find a number of related genomes that are close enough such that recognizable sequence homology in intergenic regions remains, even though a substantial fraction of nucleotides has been substituted since the common ancestor of the species. We have developed an automated pipeline that, starting from such a set of related genomes, predicts segments in intergenic regions that are under selection genome-wide. The details of this procedures, called IRUS, will be presented elsewhere. Here we briefly list the main steps:

(i) We extract the genome sequences from GenBank (18) and identify orthologous genes between all pairs of species.
(ii) We reconstruct the phylogenetic tree relating the species. We first estimate the tree topology from multiple alignments of orthologous genes. Then we determine all pairwise distances from aligned third positions in 4-fold degenerate codons. Finally we fit the pairwise distances to the tree topology to obtain the branch lengths in the tree.
(iii) We construct multiple alignments of orthologous intergenic regions using T-Coffee (19).
(iv) We scan all alignments for putative regulatory sites using a probabilistic algorithm that explicitly models the evolution of regulatory sites along the phylogenetic tree. The algorithm returns posterior probabilities for each segment to contain a regulatory site and we select a set of segments with high posterior probability.

Note that the IRUS pipeline can be applied to any set of related species for which genome sequences are available.

Reconstructing WMs from known sites and ChIP-on-chip data

For *E.coli* and *S.cerevisiae* we reconstructed a set of WMs from the known binding sites in regulonDB (1) and SCPD (2) (http://egsigma.cshl.org/jian) by an automated curation procedure using the PROCSE algorithm (20). PROCSE is a probabilistic clustering algorithm that assumes the input sequences derive from an unknown number of unknown WMs and simultaneously partitions the sites into subsets that derive from a common WM, and aligns the sequences within the subsets. For each TF we also determined the site-length that maximized the overall probability of the data. For *E.coli* this curation lead to 97 WMs for 58 different TFs and for *S.cerevisiae* to 67 WMs for 62 different TFs and complexes of multiple TFs. Second, for *S.cerevisiae* we used the extensive binding data from (15) to infer WMs using the PhyloGibbs algorithm on alignments of orthologous intergenic regions of the *Saccharomyces sensu stricto* species as described in Ref. (11). Finally we combined and hand-curated the WMs resulting from the curation of the known sites and the WMs obtained with PhyloGibbs. This led to a total of 72 high confidence WMs, most of which correspond to the binding motif of a given yeast TF, whereas a small number correspond to the binding motif of a complex of yeast TFs.

MotEvo

MotEvo is a newly developed algorithm which identifies binding sites for a set of predefined WMs by scanning multiple alignments of intergenic regions (16). MotEvo exhaustively reports putative locations of binding sites and assigns a posterior probability to each reported site. For *E.coli* we ran MotEvo with the 97 curated WMs on multiple alignments of orthologous intergenic regions from *E.coli*, *Salmonella typhi*, *Yersinia pestis KIM*, *Photorhabdus luminescens*, and *Photobacterium profundum SS9*. For this dataset MotEvo reported 6237 putative sites in the *E.coli* genome, 1162 of which have a posterior >0.5. For *S.cerevisiae* we ran MotEvo with the 72 curated WMs on the multiple alignments of orthologous intergenic regions of the *Saccharomyces sensu stricto* species. For this dataset MotEvo reported over 85 000 putative sites, of which ∼57 000 have a posterior probability >0.1 and ∼17 000 sites having a posterior probability >0.5. For each gene MotEvo was run on the multiple alignment of intergenic regions from all species for which orthologs were available. For genes for which none of the other species have an ortholog MotEvo runs on the intergenic region of the reference species only. For these cases MotEvo effectively reduces to a WM matching algorithm.

DATABASE CONTENT

Currently the SwissRegulon database contains regulatory site annotations for the following 18 organisms: *S.cerevisiae, Agrobacterium tumefaciens, Bacillus subtilis, Brucella suis, Burkholderia, Chlamydophila caviae, Corynebacterium glutamicum, Ehrlichia canis, E.coli K12, Mycobacterium tuberculosis, Neisseria meningitidis, Prochlorococcus marinus, Pseudomonas syringae, Ralstonia eutropha, Rickettsia typhi wilmington, Staphylococcus aureus, Streptococcus pneumoniae* and *Vibrio cholerae*. Our regulatory site annotations are shown in the context of the general genome annotations provided for each of these organisms. For all organisms except for yeast the genome annotations were obtained from GenBank (18). For *S.cerevisiae* the genome annotation, which is significantly more extensive, was obtained from the Saccharomyces genome database (SGD) (ftp://ftp.yeastgenome.org/yeast/; 21).

For all organisms except *E.coli* and *S.cerevisiae* the annotated regulatory sites are based on IRUS predictions only. For each site the genomic location, strand, sequence and the posterior probability as given by IRUS is recorded in the database. The number of regulatory sites predicted by IRUS varies from ∼750 sites for *E.canis* to ∼14 000 sites for *Burkholderia*. For *E.coli* and *S.cerevisiae* regulatory site annotations of MotEvo are given in addition to the IRUS predictions. For these regulatory sites the binding TF is also identified for each site. In addition, the database contains WM logos and regulons, i.e. lists of all annotated sites sorted by posterior probability for each TF. Finally, for *S.cerevisiae* the database also displays the experimentally determined binding sites from SCPD (2) and regulatory site annotations (15,17) that were downloaded from SGD. All genome-wide binding site annotations are available as flat files in gff format from the download section. For *E.coli* and *S.cerevisiae* we also provide flat files of the WMs that are used in the annotations.

DATABASE USE

The SwissRegulon database can be accessed at the address: http://www.swissregulon.unibas.ch. The database uses the Generic genome browser (GBrowse) (22) as an engine and is fully compatible with the original GBrowse. For a detailed description of GBrowse usage and features please see the original manual at the developers page (http://www.gmod.org/gbrowse). Briefly, the genome browser graphically displays a section of the genome and all features annotated on it. The user can zoom in and out and scroll through the genome and click on features to obtain more detailed information.

Users can specify a genome segment for displaying, e.g. chrII:600..1000, or query the database by entering a keyword including wild card characters, e.g. SKO*. This query will return a list of matches to the search term. For example, to find all annotated binding sites for the transcription factor RAP1 one would query the database for RAP1* (note that, beyond the binding sites this query would also return the RAP1 gene). By clicking on one of the sites in the list the user will see the section of the genome where the site occurs.

Annotated regulatory sites are displayed as rectangular boxes with an arrow inside showing the strand of the site. The posterior probability assigned to the site is represented by the intensity of the box's color. That is, the higher the posterior probability, the more intense is the color of the box. Every box is labeled by an identifier which is either the name of the TF that binds the site or a unique identifier if the site has not been assigned to any known TF.

An example screen shot is shown in Figure 1. Placing the cursor on the box brings up a pop-up legend with the sequence of the site and its posterior probability. Clicking on a binding site box links to a page with detailed information about the site. For binding sites assigned to a TF this information includes the 'regulon list' of all sites for the same TF, and a logo of its WM. The regulon list shows for each site in the regulon the name(s) of the upstream gene(s) it regulates, the genomic coordinates of the intergenic region in which it occurs, the genomic coordinates of the site, and the posterior probability of the site. For convenient browsing the user can filter out sites according to their posterior probability. Filters are accessible under the 'Results and Analysis' pop-up menu.

COMPARISON WITH EXISTING RESOURCES

There are a number of databases that collect known TF binding sites from the literature. Most of these focus on regulatory sites from a single organism, e.g. RegulonDB for *E.coli* (23), SCPD for *S.cerevisiae* (2) and the more recent regulatory site annotations based on ChIP-on-chip data (15,17), DBTBS for *B.subtilis* (24), AGRIS for *Arabidopsis thaliana* (25), and the DNase I footprint database for *Drosophila melanogaster* (26). Similarly, there are a number of databases that contain known binding sites in vertebrate genomes (27–29). The well-known commercial TRANSFAC database (30) probably contains the

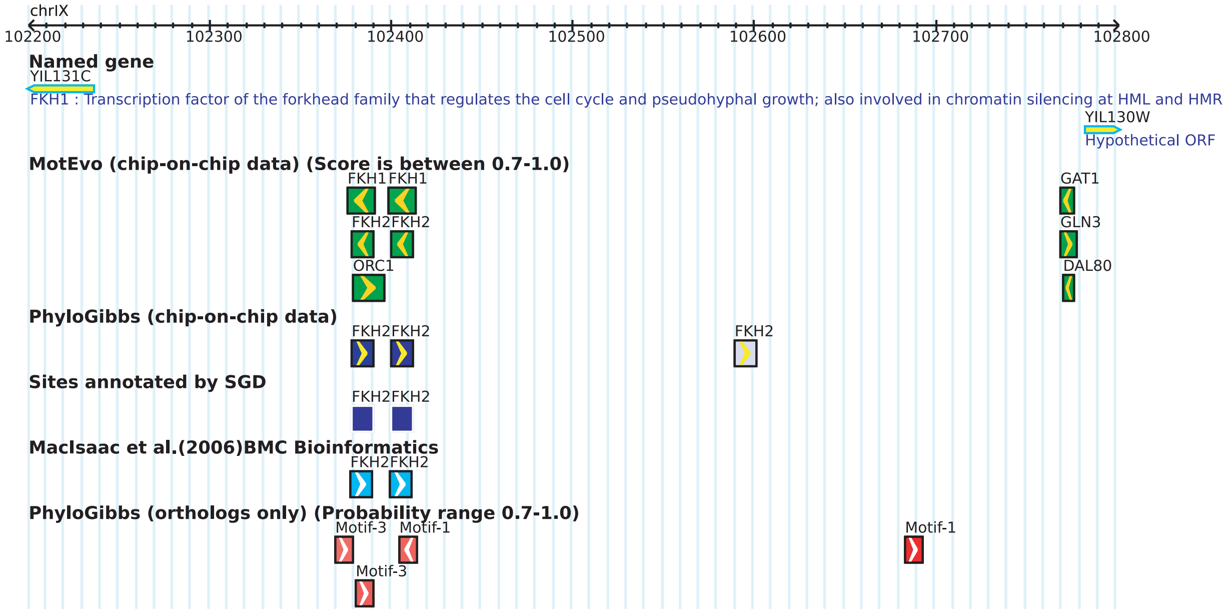
FKH1 : Transcription factor of the forkhead family that regulates the cell cycle and pseudohyphal growth; also involved in chromatin silencing at HML and HMR

Figure 1. Screen shot from the SwissRegulon database for the intergenic region between the genes FKH1 and YIL130W in *S.cerevisiae*. Only sites with posterior probability 0.7 or larger are shown.

largest collection of experimentally determined binding sites from multiple organisms, mainly from eukaryotic organisms. Finally, the PRODORIC database (31) focuses on prokaryotic genomes and contains collections of known binding sites from a number of bacteria, with *E.coli*, *B.subtilis*, and *Pseudomonas aeruginosa* represented by a substantial number of sites. Most of these databases also contain WMs for the TFs for which known sites are available (32). Some of the databases also offer the possibility of scanning intergenic regions with these WMs, and in some cases to filter the resulting sites for conservation in related species. Additionally, databases and web servers have been made available that show the results of 'phylogenetic footprinting' methods (33–35), i.e. that display conservation profiles for particular sets of related genomes.

Over the last years we have developed a number of probabilistic methods (11,16,20) for rigorously combining information from known binding sites and ChIP-on-chip data with motif finding methods, and phylogenetic footprinting. By applying these methods we obtain genome-wide regulatory site annotations across different genomes using a unified methodology, which rigorously assigns quality estimates, i.e. posterior probabilities, to all predicted sites. The main aim of SwissRegulon is to make these regulatory site annotations available across as many genomes as possible, both prokaryotic and eukaryotic. In addition we make all the annotations available using a common GBrowse genome browser interface that shows the binding sites in the context of other features annotated on the genome. Through this user-friendly graphical interface the SwissRegulon resource will be useful for people researching regulatory mechanisms both experimentally and computationally.

FUTURE DEVELOPMENTS

In the near future SwissRegulon will significantly expand the number of organisms represented, especially bacterial ones. For the bacteria for which significant collections of known sites exist, e.g. *B.subtilis* and *P.aeruginosa*, we will include these into the predictions as currently done for *E.coli*. Eventually we also intend to include comprehensive regulatory site annotations for higher eukaryotes, i.e. vertebrate genomes, flies and worms.

Second, we are intending to incorporate ChIP-on-chip data in the SwissRegulon database in the near future. The combination between the binding site annotations and condition-specific ChIP-on-chip data will give insight into the conditions under which different sites are bound by their TFs.

Third, currently binding sites are shown on a per genome basis even though site conservation across related organisms is used in the predictions. In the future we intend to provide explicit information about conservation for each binding site and to link each binding site to the orthologous binding sites in the related genomes.

Finally, we have recently implemented a web server (http://www.phylogibbs.unibas.ch) for running the PhyloGibbs motif and regulatory site finding algorithm (11). In the future we intend to integrate these two resources. This will allow users to run PhyloGibbs on input data that was selected in the genome browser, and to see the results in the context of the existing regulatory site annotations.

ACKNOWLEDGEMENTS

The research in this study was supported by SNF grant 3152A0-105972. Funding to pay the Open Access publication charges for this article was provided by the Biozentrum, University of Basel.

Conflict of interest statement. None declared.

REFERENCES

1. Salgado,H., Santos-Zavaleta,A., Gama-Castro,S., Millan-Zarate,D., Blattner,F. and Collado-Vides,J. (2000) RegulonDB (version 3.0): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Res*., **28**, 65–7.
2. Zhu,J. and Zhang,M.Q. (1999) SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics*, **15**, 607–611.
3. Stormo,G.D. and Hartzell,G.W. (1989) Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl. Acad. Sci. USA*, **86**, 1183–1187.
4. Bailey,T. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol*., **2**, 28–36.
5. Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.
6. McCue,L.A., Thompson,W., Carmack,C.S., Ryan,M.P., Liu,J.S., Derbyshire,V. and Lawrence,C.E. (2001) Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res*., **29**, 774–782.
7. Rajewsky,N., Socci,N.D., Zapotocky,M. and Siggia,E.D. (2002) The evolution of DNA regulatory regions for proteo-gamma bacteria by interspecies comparisons. *Genome Res*., **12**, 298–308.
8. Wang,T. and Stormo,G. (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, **19**, 2369–2380.
9. Moses,A.M., Chiang,D.Y. and Eisen,M.B. (2004) Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. *Pac. Symp. Biocomput*., **2004**, 324–335.
10. Sinha,S., Blanchette,M. and Tompa,M. (2004) PhyME: A probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics*, **5**, 170.
11. Siddharthan,R., Siggia,E.D. and van Nimwegen,E. (2005) Phylogibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput. Biol*., **1**, e67.
12. Cliften,P., Sudarsanam,P., Desikan,A., Fulton,L., Fulton,B., Majors,J., Waterston,R., Cohen,B.A. and Johnston,M. (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science*, **301**, 71–76.
13. Kellis,M., Patterson,N., Endrizzi,M., Birren,B. and Lander,E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.
14. Lee,T.I., Rinaldi,N.J., Robert,F., Odom,D.T., Bar-Jospeh,Z., Gerber,G.K., Hannett,N.M., Harbison,C.T., Thompson,C.M., Simon,I. *et al*. (2002) Transcription regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.
15. 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., Yoo,J. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.
16. Erb,I. and van Nimwegen,E. (2006) Statistical features of yeast's transcriptional regulatory code. *IEEE Proceedings of the first International Conference on Computational Systems Biology (ICCSB)*, 111–118.
17. Macisaac,K.D., Wang,T., Gordon,B.D., Gifford,D.K., Stormo,G.D. and Fraenkel,E. (2006) An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics*, **7**, 113.
18. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2006) GenBank. *Nucleic Acids res*., **34**, D16–D20.
19. Notredame,C., Higgins,D. and Heringa,J. (2000) T-Coffee: a novel method for multiple sequence alignments. *J. Mol. Biol*., **302**, 205–217.

20. van Nimwegen,E., Zavolan,M., Rajewsky,N. and Siggia,E.D. (2002) Probabilistic clustering of sequences: inferring new bacterial regulons by comparative genomics. *Proc. Natl Acad. Sci. USA*, **99**, 7323–7328.
21. Hong,E.L., Balakrishnan,R., Christie,K.R., Dwight,S.S., Engel,S.R., Fisk,D.G., Hirschman,J.E., Livstone,M.S., Nash,R., Park,J. *et al*. (2006) Saccharomyces genome database.
22. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al*. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res*., **12**, 1599–1610.
23. Salgado,H., Gama-Castro,S., Peralta-Gil,M., Diaz-Peredo,E., Sanchez-Solano,F., Santos-Zavaleta,A., Martinez-Flores,I., Jimenez-Jacinto,V., Bonavides-Martinez,C., Segura-Salazar,J. *et al*. (2006) RegulonDB (version 5.0): *Escherichia coli* k-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res*., **34**, D394–D397.
24. Makita,Y., Nakao,M., Ogasawara,N. and Nakai,K. (2004) DBTBS: database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics. *Nucleic. Acids Res*., **32**, D75–D77.
25. Davuluri,R.V., Sun,H., Palaniswamy,S.K., Matthews,N., Molina,C., Kurtz,M. and Grotewold,E. (2003) AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. *BMC Bioinformatics*, **4**, 25.
26. Bergman,C.M., Carlson,J.W. and Celniker,S.E. (2004) Drosphila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, *Drosophila melanogaster*. *Bioinformatics*, **21**, 1747–1749.
27. Blanco,E., Farre,D., Alba,M.M., Messeguer,X. and Guigo,R. (2006) ABS: a database of Annotated regulatory binding sites from orthologous promoters. *Nucleic Acids Res*., **34**, D63–D67.
28. Zhao,F., Xuan,Z., Liu,L. and Zhang,M.Q. (2005) Tred: a transcriptional regulatory element database and a platform for in silico gene regulation studies. *Nucleic Acids Res*., **33**, D103–D107.
29. Gosh,D. (2000) Object-Oriented Transcription Factors Database (ooTFD). *Nucleic Acids Res*., **28**, 308–310.
30. Wingender,E., Dietze,P., Karas,H. and Knüppel,R. (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res*., **24**, 238–241.
31. Münch,R., Hiller,K., Barg,H., Heldt,D., Linz,S., Wingender,E. and Jahn,D. (2003) PRODORIC: prokaryotic database of gene regulation. *Nucleic Acids Res*., **31**, 266–269.
32. Vlieghe,D., Sandelin,A., Bleser,P.J.D., Vleminckx,K., Wasserman,W.W., van Roy,F. and Lenhard,B. (2006) A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res*., **34**, D95–D97.
33. Frazer,K.A., Pachter,L., Poliakov,A., Rubin,E.M. and Dubchak,I. (2004) VISTA: computational tools for comparative genomics. *Nucleic Acids Res*., **32**, W273–W279.
34. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al*. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*., **15**, 1034–1050.
35. Dieterich,C., Wang,H., Rateitschak,K., Luz,H. and Vingron,M. (2003) CORG: a database for COmparative Regulatory Genomics. *Nucleic Acids Res*., **31**, 55–57.