BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 23 2006, pages 2940–2944
doi:10.1093/bioinformatics/btl509

Databases and ontologies

CREMOFAC—a database of chromatin remodeling factors

Agrawal Shipra[1,†], Kumar Chetan[1] and M. R. S. Rao[1,2,*]

[1]Molecular Biology and Genetics Unit, Jawaharlal Nehru Centre for Advanced Scientific Research (JNCASR), Jakkur, Bangalore 560064, India and [2]Department of Biochemistry, Indian Institute of Science, Bangalore 560012, India

Received on July 20, 2006; revised on September 13, 2006; accepted on September 29, 2006
Advance Access publication October 4, 2006
Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: Chromatin-remodeling is an important event in the eukaryotic nucleus rendering nucleosomal DNA accessible for various transaction processes. Remodeling Factors facilitate the dynamic nature of chromatin through participation of the collective action of (i) ATP and (ii) Non-ATP dependent factors. Considering the importance of these factors in eukaryotes, we have developed, CREMOFAC, a dedicated and frequently updated web-database for chromatin-remodeling factors.

Results: The database harbors factors from 49 different organisms reported in literature and facilitates a comprehensive search for them. In addition, it also provides in-depth information for the factors reported in the three widely studied mammals namely, human, mouse and rat. Further, information on literature, pathways and phylogenetic relationships has also been covered. The development of CREMOFAC as a central repository for chromatin-remodeling factors and the absence of such a pre-existing database heighten its utility thus making its presence indispensable.

Availability: http://www.jncasr.ac.in/cremofac/

Contact: mrsrao@jncasr.ac.in

1 INTRODUCTION

The eukaryotic genome is organized as a highly compact DNA–protein complex termed 'chromatin'. Its organization as chromatin serves to compact the DNA in a small volume of the nucleus. It is well established that the DNA transaction processes, such as gene expression, recombination, duplication or repair require the flexibility of the chromatin structure. This re-arrangement is largely driven by the action of highly conserved molecular machines broadly termed as remodeling enzyme complexes (Becker, 2005). A family of remodeling complexes consists of related enzymes commonly sharing an ATPase subunit that remodels chromatin by utilizing energy of ATP hydrolysis (Vignali *et al*., 2000). This family can be further subdivided into three families based on their biochemical functions and presence of the protein motifs outside their ATPase subunits: (i) the SWI/SNF family (Peterson and Workman, 2000), (ii) ISWI and (iii) Mi-2/CHD families (Boyer *et al*., 2000). Most of the members of SWI/SNF family play an important role in the activation of transcription, whereas the members belonging to ISWI and Mi-2/CHD families are primarily dedicated to transcriptional repression pathways (Deuring *et al*., 2000). Besides their transcriptional role, the members of ISWI family have also been reported to participate in chromatin assembly and also facilitate other chromatin based diverse processes like recombination and DNA repair (Peterson, 2002). Additionally, chromatin-remodeling complexes require concerted action with HATs and HDACs in order to express and repress the genes, respectively.

More recent studies have proposed the involvement of the remodeling enzymes in altering the topology of nucleosomal DNA, DNA tracking activities (Saha *et al*., 2005), rotation of DNA along its long axis or formation of DNA bulges or small loops (Peterson, 2002). Although the remodeling of chromatin by these enzymes controls the reprogramming of chromatin, an inappropriate remodeling may lead to enhanced rate of DNA recombination, disruption of chromosome condensation, or deregulation of gene expression (Roberts and Orkin, 2004). Such conditions probably arise due to the improper recruitment of these enzymes, which in turn may lead to certain pathological conditions like cancer (Roberts and Orkin, 2004). Furthermore, studies on remodeling complexes have led to the identification of numerous complexes in organisms and a surfeit of data on their structural and functional aspects has got accumulated. With an aim to provide convenient access to this large volume of data as well as be able to query it through various parameters, the present CREMOFAC database has been developed.

CREMOFAC, a collection of such remodeling factors reported in mammals and other higher eukaryotes, houses chromatin-remodeling factors from 49 different organisms reported in literature and provides an interface to enable a quick and efficient search. With the aid of a highly user-friendly interface, families of ATP-dependent remodelers (ISWI, Swi/Snf2, CHD or Mi-2) and Non-ATP dependent remodelers (HDACs and HATs) can be queried comprehensively through various arguments, such as factor names, organism, family names and homology. For every such information retrieved, the users can effortlessly download it. The homology search enables a user to identify the homologous sequences across the whole set of factors in various organisms. Further, the 'Phylogeny Trees' feature provides a graphical representation of the phylogenetic relationship for a factor reported in different organisms. By providing links to other resourceful databases, Pubmed literature and useful analysis tools, it provides a platform from where information on chromatin-remodeling complexes and pathways can be conveniently retrieved. Therefore, the availability of such an extensive and comprehensive database

*To whom correspondence should be addressed.
†Present address: Institute of Bioinformatics and Applied Biotechnology, ITPL, Bangalore, India
The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

 © The Author 2006. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

Downloaded from bioinformatics.oxfordjournals.org by guest on May 19, 2011

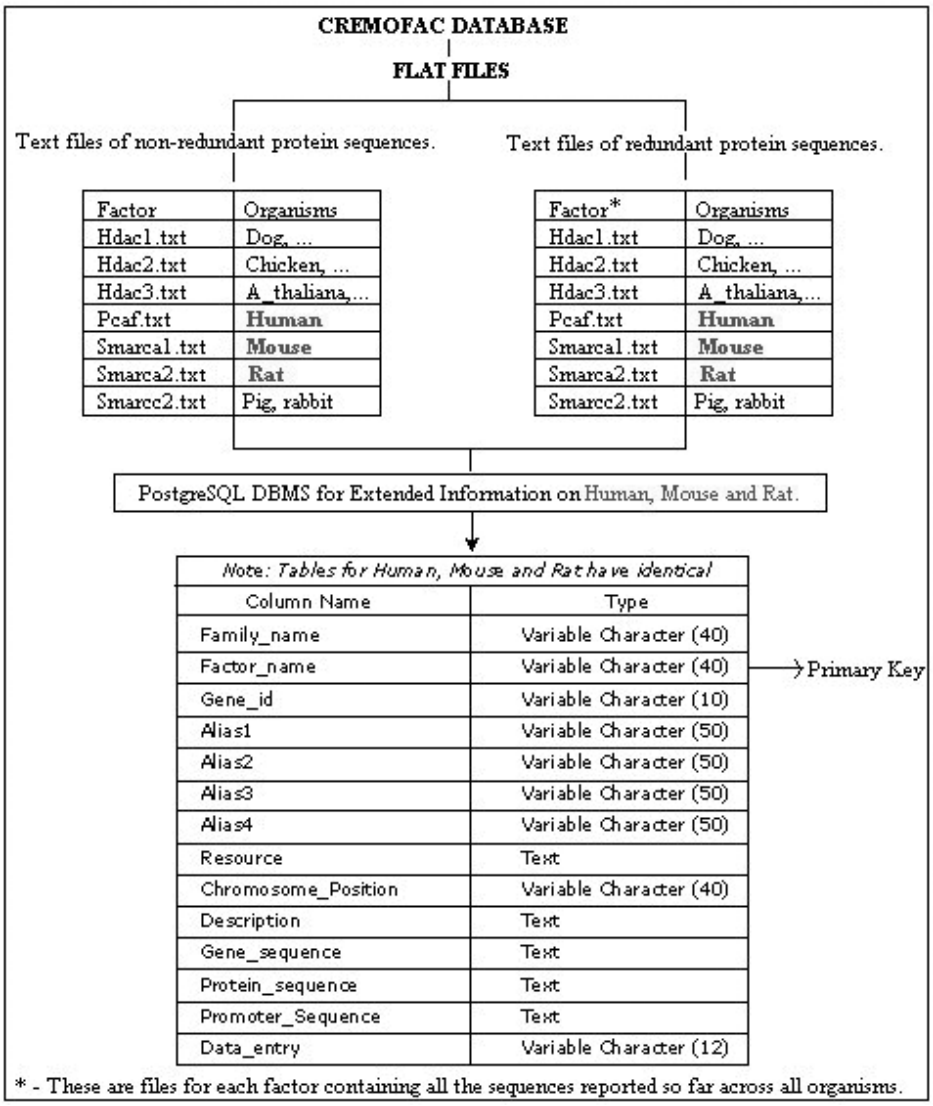

Fig. 1. A block diagram of CREMOFAC database structure and design. The database comprises of a collection of flat files for every factor (redundant/non-redundant protein sequences), and a PostgreSQL relational database management system that stores the extended information for human, mouse and rat remodeling factors in the form of three different identical tables as shown in the figure.

would be of interest for a wide range of researchers working on chromatin structure and function. The database will be updated with further enhanced features and new factors as and when they are identified.

2 METHODS

2.1 Procurement and arrangement of sequences

The orthologous and paralogous protein sequences of remodeling factors across different organisms were obtained using the Gap BLAST and PSI-BLAST programs against non-redundant database at default parameter values. For further clarification, the definition of orthologous, paralogous and homologous sequences is as follows. Orthologous factors constitute the set of unique factors found in multiple organisms. Paralogous factors defined for an organism comprise of multiple occurrences of a single factor found at same/different chromosome(s), and Homologous factors of a

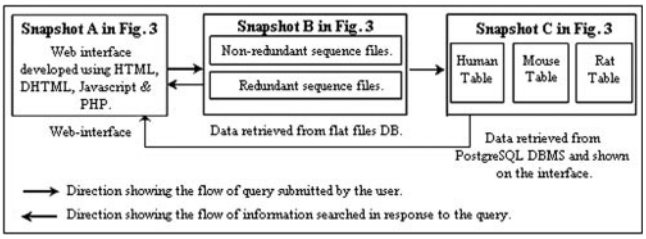

Fig. 2. A bird's eye view of the direction of flow of information when the query in Figure 3 is submitted.

remodeling factor found in an organism is the collection of highly similar sequences, which were also predicted to have the same functional domains in the same organism. The sequences thus obtained for each factor were filtered to remove similar matches as factors are highly conserved within the

Downloaded from bioinformatics.oxfordjournals.org by guest on May 19, 2011

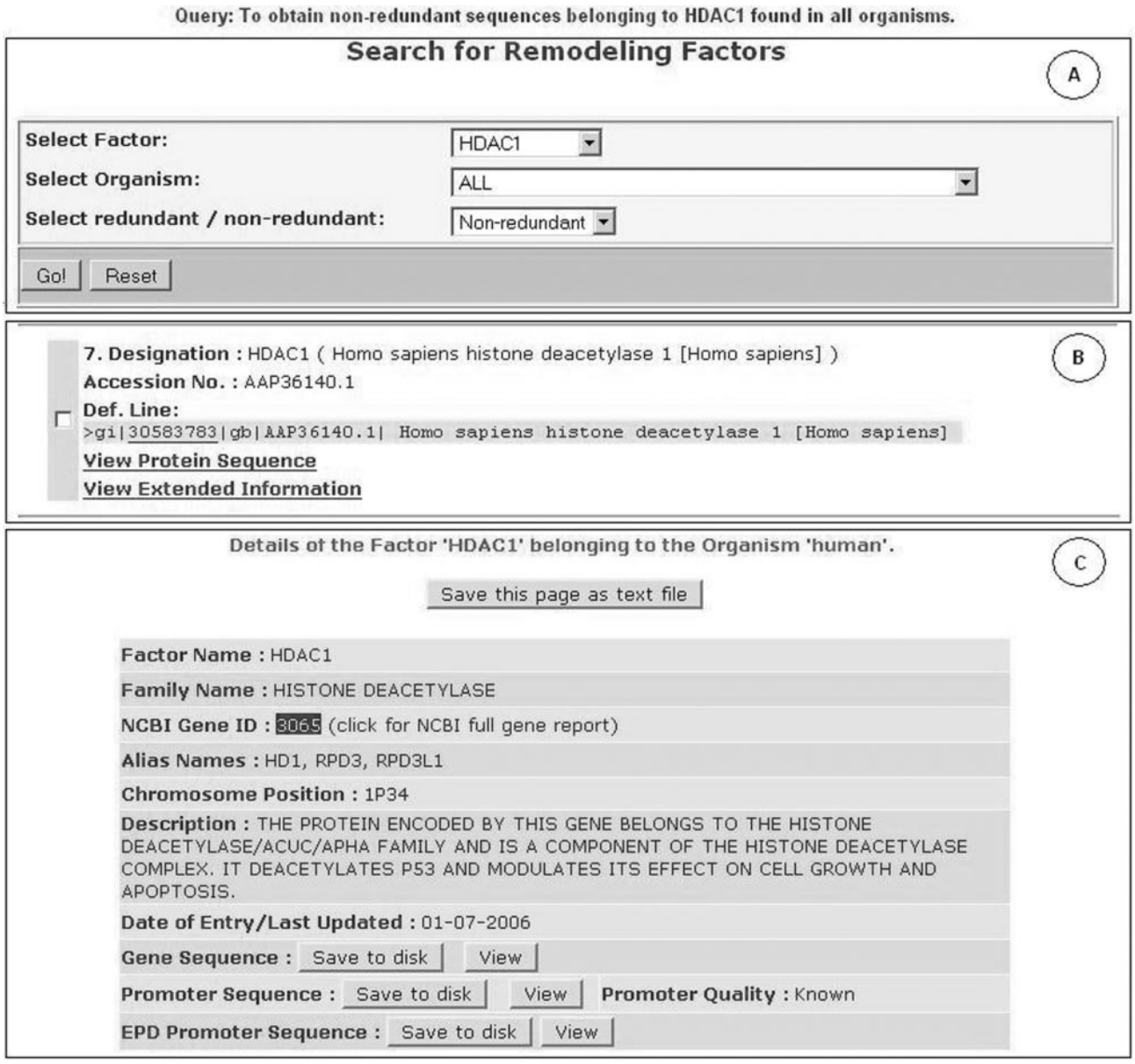

Fig. 3. (**A**) 'Simple Search' forms to take input from the user to frame the query. (**B**) Result of the search, showing 'link to extended information' for Human. (**C**) On clicking 'view extended information' for the human sequence in B, the extended information is retrieved from the PostgreSQL database and shown as C.

family. After having accomplished this task, it was essential to have a well-clustered data, to further enhance the data quality and data comprehensibility. To achieve this, a computational program was written and executed to sort the sequences in alphabetical order of the name of the organism in which they occur. After this re-arrangement of sequences, two sets of files were made for every factor, namely redundant and non-redundant sets. The redundant sequence set for each of the factors includes all identified orthologous/paralogous/homologous factor information. The non-redundant set stores information on the longest protein sequence(s) of each factor in the corresponding organisms comprising of conserved and flanking region(s). The idea behind this segregation was to filter and provide the full-length protein sequences from truncated or fragment sequences. Henceforth, the user can obtain information for the non-redundant factors in a selected organism.

2.2 Sources and content of 'extended information'

The 'extended information' on the chromatin-remodeling factors found in human, mouse and rat has been extracted from NCBI (http://www.ncbi.nlm.nih.gov/), Ensembl (http://www.ensembl.org/index.html), Mouse (http://www.informatics.jax.org/) and Rat Genome Database (http://rgd.mcw.edu/). The data has been categorized into classes or families depending upon the function of remodeling factors. These classes are ISWI, Swi/Snf2, CHD or Mi-2, bromodomain chromatin modifiers (BRD and BAZ enzymes) and other ARID proteins capable of modifying chromatin. In addition, data on HDACs and HATs has also been compiled in CREMOFAC. The known promoter sequences (1000 bp upstream and 100 bp downstream of the transcription start site) were downloaded from TRED (http://rulai.cshl.edu/cgi-bin/TRED/tred.cgi?process=home) database whereas the new

Downloaded from bioinformatics.oxfordjournals.org by guest on May 19, 2011

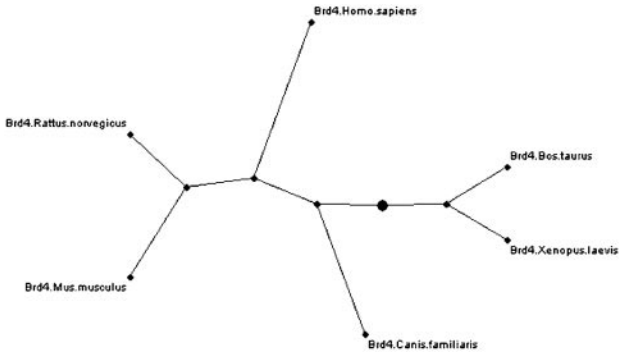

Fig. 4. Snapshot of a phylogenetic tree as displayed by CREMOFAC.

promoters were predicted using transcription start site and first exon information from DBTSS (http://dbtss.hgc.jp/). The promoter information has been supplemented with the corresponding promoter data available in Eukaryotic Promoter Database (EPD: http://www.epd.isb-sib.ch/). Annotation of the protein domains and the prediction phylogenetic tree were done using SMART (http://smart.embl-heidelberg.de/) and Phylodraw (http://pearl.cs.pusan.ac.kr/phylodraw/) programs, respectively.

2.3 Web-database system, platforms and specifications

The CREMOFAC database is a fusion of flat files and PostgreSQL, a relational database management system (RDBMS) operating on UNIX platform. The RDBMS stores the extended data for human, mouse and rat, whereas the flat files house orthologous protein sequences for remodeling factors found in different organisms. The databases as well as the website are hosted at the JNCASR web server (http://www.jncasr.ac.in), which operates on a Linux operating system. The web interface has been developed using PHP, HTML, DHTML and JavaScript languages. The database makes use of a well-organized directory structure of files and relational tables to service the information as per the query. A schematic diagram of the database structure is further illustrated in Figure 1.

3 IMPLEMENTATION

3.1 Database structure and design

The database resembles two-tier architecture, the first tier being that of the flat files while the second being the RDBMS database comprising of extended information for human, mouse and rat. Owing to such a structure, the results are presented in a step-wise fashion so that user can either view the information on all orthologous remodeling factors, or extended information for the factors found in human, mouse or rat. The flat file database is made up of two sets of files, which comprise of redundant and non-redundant factor sequences across all reported organisms. The PostgreSQL database is made up of three tables, each for human, mouse and rat having identical structure. Figures 1–3, further exemplify the database structure. Figure 1 represents a schematic diagram of the CREMOFAC database structure and design. Figure 2 gives a bird's eye view of the interaction between various components and how CREMOFAC furnishes information when a user, as in Figure 3, submits a query. Figure 3 shows snapshots of the database and systematic flow of information when a query is submitted.

4 RESULTS AND DISCUSSION

4.1 Data statistics of CREMOFAC database

The CREMOFAC database comprises of 64 types of remodeling factors that have been reported across 49 different organisms till date. In totality, the database houses 1725 remodeling factors in the redundant dataset and 720 factors in the non-redundant dataset. A comprehensive list of the organisms as well as the number of factors found in them is available at the CREMOFAC website under the section, 'Database Statistics'. A list of all organisms found to possess a particular factor has also been provided.

4.2 Access to CREMOFAC and salient features

The CREMOFAC web site and database has been designed to cater to the needs of the researchers working in the area of transcriptional research as well as the neophyte students working on chromatin-remodeling factors. One can query the database using a wide range of search parameters. The salient features of the database are listed as follows.

(1) It provides gene, protein, promoter and isoform protein sequences, protein domain images as well as other details, such as chromosome position, family name, direct link to NCBI gene report page and a brief description of all the factors found in the three widely studied mammals, human, mouse and rat.

(2) It provides various search parameters to query the database thoroughly, such as search by name of family, factor or organism, search within ATP dependent factors or Non-ATP dependent factors, as well as search for redundant or non-redundant sequences. It also allows the user to search for homologous protein sequences. Through this feature, the location of conserved region(s) in sequences may also be identified.

(3) The 'Phylogeny Trees' feature can be of significant importance to biologists as it shows a graphical representation of the evolutionary interrelationships among various species reported in literature. Moreover, downloadable Phylodraw files for each of these trees have also been provided to allow for in-depth analysis of these trees and its components. Figure 4 represents a snap shot of a phylogenetic tree as displayed by CREMOFAC.

(4) The 'Remodeling Factors' section in CREMOFAC shows an introductory note on remodeling factors, different families as well as schematic view of the chromatin-remodeling function. The same section displays chromatin-remodeling pathway diagrams obtained from Biocarta database (http://www.biocarta.com). The user can further see the details of pathway enzymes by clicking on them in the figure.

(5) The software allows the user to select individual, multiple or all sequences retrieved as results and either download or view them as text files. Wherever possible, options to save or view these sequences have been provided for their fast and easy procurement. Links to PubMed have also been provided to guide the user to the key publications that have contributed to this field. It addition, it provides links to softwares pertaining to the analysis of remodeling factors.

Downloaded from bioinformatics.oxfordjournals.org by guest on May 19, 2011

(6) The CREMOFAC database contents can be downloaded easily from the 'Download Database' section. Users can obtain the entire collection of sequences with a single mouse click as well as download phylogeny trees and protein domain images.

ACKNOWLEDGEMENTS

The authors are also grateful to Ms Surbhi Dhar, Dr Jayashree Ladha, Mr Pradeepa M.M. and Ms G.Gayathri for their useful suggestions. The authors are earnestly thankful to the reviewers of the first version of CREMOFAC for their constructive suggestions in improving the database, thereby making it more useful and valuable. Support from JNCASR, Bangalore for computational facilities and financial assistance is greatly acknowledged.

Conflict of Interest: none declared.

REFERENCES

Boyer,L.A. *et al*. (2000) Functional delineation of three groups of the ATP-dependent family of chromatin remodeling enzymes. *J. Biol. Chem.*, **275**, 18864–18870.

Deuring,R. *et al*. (2000) The ISWI chromatin-remodeling protein is required for gene expression and the maintenance of higher order chromatin structure *in vivo*. *Mol. Cell*, **5**, 355–364.

Becker,P.B. (2005) Nucleosome remodelers on track. *Nat. Struct. Mol. Biol.*, **12**, 732–733.

Peterson,C.L. (2002) Chromatin remodeling enzymes: taming the machines. *EMBO Rep.*, **3**, 319–322.

Peterson,C.L. and Workman,J.L. (2000) Promoter targeting and chromatin remodeling by the SWI/SNF complex. *Curr. Opin. Genet. Dev.*, **10**, 187–192.

Roberts,C.W. and Orkin,S.H. (2004) The SWI/SNF complex--chromatin and cancer. *Nat. Rev. Cancer*, **4**, 133–142.

Saha,A. *et al*. (2005) Chromatin remodeling through directional DNA translocation from an internal nucleosomal site. *Nat. Struct. Mol. Biol.*, **12**, 747–55.

Vignali,M. *et al*. (2000) ATP-dependent chromatin-remodeling complexes. *Mol. Cell. Biol.*, **20**, 1899–1910.

Downloaded from bioinformatics.oxfordjournals.org by guest on May 19, 2011