336–337 *Nucleic Acids Research, 1999, Vol. 27, No. 1*

The Homeodomain Resource: sequences, structures and genomic information

Sharmila Banerjee-Basu, Erik S. Ferlanti, Joseph F. Ryan and Andreas D. Baxevanis*

Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Building 49, Room 2C-76, Bethesda, MD 20892-4431, USA

Received August 31, 1998; Revised September 7, 1998; Accepted September 28, 1998

ABSTRACT

The Homeodomain Resource is a comprehensive collection of sequence, structure and genomic information on the homeodomain protein family. Available through the Resource are both full-length and domain-only sequence data, as well as X-ray and NMR structural data for proteins and protein–DNA complexes. Also available is information on human genetic diseases and disorders in which proteins from the homeodomain family play an important role; genomic information includes relevant gene symbols, cytogenetic map locations, and specific mutation data. Search engines are provided to allow users to easily query the component databases and assemble specialized data sets. The Homeodomain Resource is available through the World Wide Web at http://genome.nhgri.nih.gov/homeodomain

INTRODUCTION

The homeodomain is a highly-conserved structural motif of ~60 amino acid residues that is found in many eukaryotic transcription factors. Homeodomain proteins play a fundamental role in diverse developmental processes, including the specification of body plan, pattern formation, and the determination of cell fate (1). X-ray crystallographic and NMR spectroscopic studies on several members of this family revealed that these proteins contain three helical regions folded into a compact, globular structure, with an N-terminal extension (2–7). Helices I and II lie parallel to each other and across from the third helix, which is also known as the recognition helix. This third helix, in conjunction with the N-terminal arm, confers the DNA-binding specificity of individual homeodomain proteins. The regulatory function of a homeodomain protein is based upon its specific interactions with the transcriptional control region of a target gene.

The homeodomain has been identified in a broad spectrum of organisms, ranging from yeast to *Drosophila* to humans (see ref. 8 for a review}. Furthermore, a number of homeodomain proteins have been implicated in human genetic and genomic disorders, such as aniridia (OMIM 106210), cone-rod retinal dystrophy (OMIM 602225), and Waardenburg syndrome (OMIM 193500). As an extension of genomically-based structural studies currently being performed, we have assembled a comprehensive collection of homeodomain resources. This collection is intended to be a central source of sequence, structural and genomic information on this important class of developmental regulatory proteins. This database will be continuously updated, based on both public database searches and on submissions from the homeodomain community.

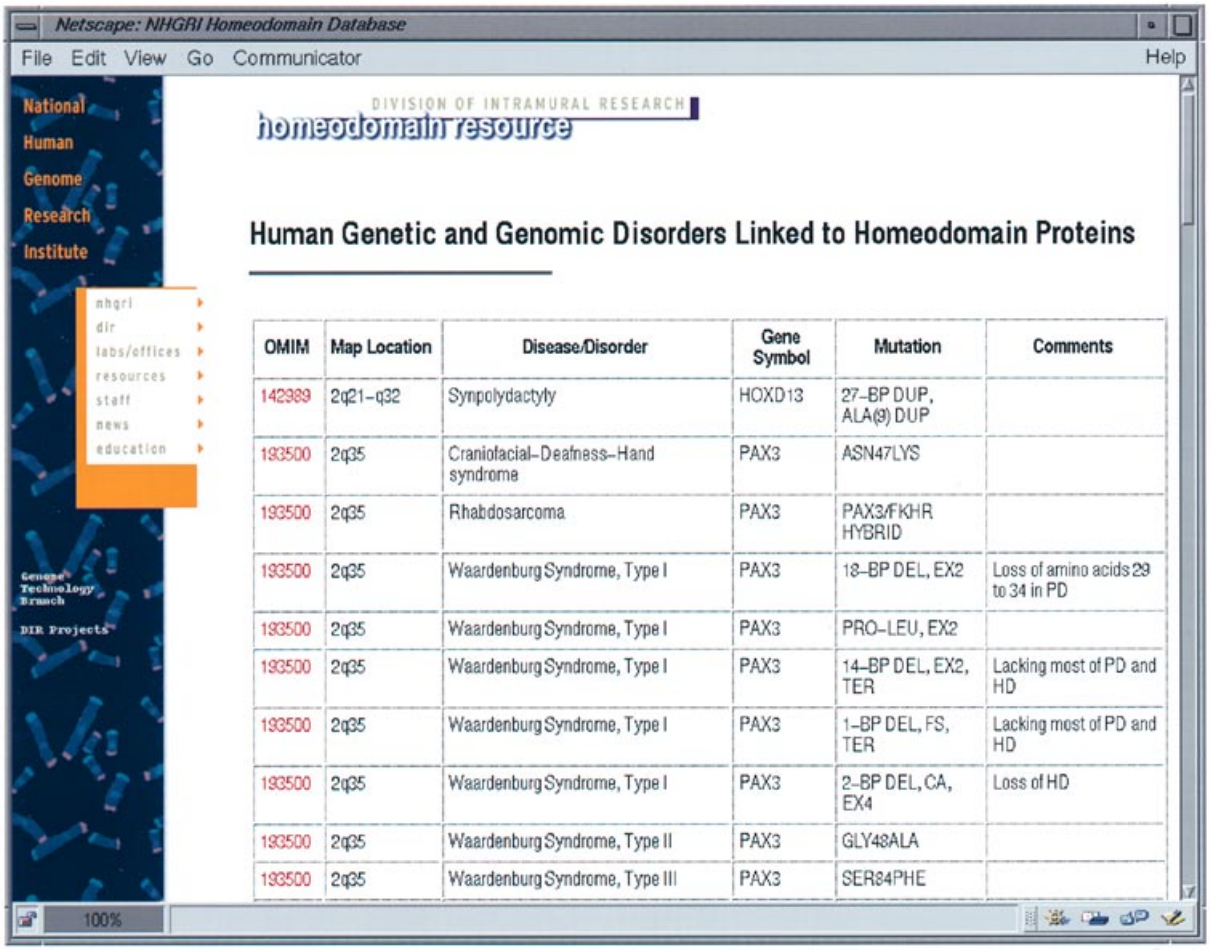
DATABASE CONTENT

Protein sequence data represents a compilation of entries from SWISS-PROT (9) as of late August, 1998. Complete sequence data is available in FASTA format, with links to NCBI Entrez (10). A second FASTA-format sequence set, showing only the homeodomain portion of the complete sequence, is also available. A simple search engine is available for searching the sequence data, using the SWISS-PROT ID, GenBank accession number, protein name, organism name, gene name, or sequence pattern as the query; hits are returned in FASTA format. A link is also provided to NCBI Entrez for real-time queries against the protein databases; results of this search will yield a complete but unrefined set of all homeodomain sequences.

A large number of X-ray and NMR structures have been determined for various homeodomain proteins, including Antennapedia (11–14) and engrailed (2,15,16). A number of these proteins have also been co-crystallized with DNA so that structural comparisons can be made between the free and bound proteins. A list of these structures is available through the Homeodomain Resource, with links to both the Protein Data Bank (http://www.pdb.bnl.gov) and MMDB (17). For structural solutions of a protein–DNA complex, the DNA ligand is listed, with an indication of which nucleotides are involved in forming the protein–DNA complexes. From the MMDB entry, users can view the structure itself using Cn3D, a molecular viewing application that is bundled with Network Entrez and can be downloaded by following hyperlinks on any structure entry page (18).

The initial set of genomic information was compiled from both the literature and from the Online Mendelian Inheritance in Man database at NCBI (http://www.ncbi.nlm.nih.gov/Omim/). The available information includes gene symbols, protein names, disease names, cytogenetic map locations, and relevant mutation data possibly leading to a human genetic disorder (Fig. 1). This

*To whom correspondence should be addressed. Tel: +1 301 496 8570; Fax: +1 301 402 6858; Email: andy@nhgri.nih.gov



OMIM	Map Location	Disease/Disorder	Gene Symbol	Mutation	Comments
142989	2q21–q32	Sympolydactyly	HOXD13	27–BP DUP, ALA(9) DUP	
193500	2q35	Craniofacial–Deafness–Hand syndrome	PAX3	ASN47LYS	
193500	2q35	Rhabdosarcoma	PAX3	PAX3/FKHR HYBRID	
193500	2q35	Waardenburg Syndrome, Type I	PAX3	18–BP DEL, EX2	Loss of amino acids 29 to 34 in PD
193500	2q35	Waardenburg Syndrome, Type I	PAX3	PRO–LEU, EX2	
193500	2q35	Waardenburg Syndrome, Type I	PAX3	14–BP DEL, EX2, TER	Lacking most of PD and HD
193500	2q35	Waardenburg Syndrome, Type I	PAX3	1–BP DEL, FS, TER	Lacking most of PD and HD
193500	2q35	Waardenburg Syndrome, Type I	PAX3	2–BP DEL, CA, EX4	Loss of HD
193500	2q35	Waardenburg Syndrome, Type II	PAX3	GLY48ALA	
193500	2q35	Waardenburg Syndrome, Type III	PAX3	SER84PHE	

Figure 1. Example of tabular genomic and genetic data available through the Homeodomain Resource, in this case sorted by cytogenetic map location. Hyperlinks are provided to Online Mendelian Inheritance in Man. A Web front-end to the Sybase database is also available for both searching the mutation data and generating custom tables.

information is available in a number of tabular formats. In addition, a Web front-end to a Sybase database has been developed for the homeodomain community to contribute and query mutation information that may not appear in OMIM or elsewhere. Users are encouraged to inform the authors of any updates or corrections to database information. Unpublished information will be held as confidential upon request.

REFERENCES

1 Gehring,W., Affolter,M. and Bürglin,T. (1994) *Annu. Rev. Biochem.*, **63**, 487–526.
2 Kissinger,C., Liu,B., Martin-Blanco,E., Kornberg,T. and Pabo,C. (1990) *Cell*, **63**, 579–590.
3 Wolberger,C., Vershon,A., Liu,B., Johnson,A. and Pabo,C. (1991) *Cell*, **67**, 517–528.
4 Wilson,D.S., Guenther,B., Desplan,C. and Kuriyan,J. (1995) *Cell*, **82**, 709–719.
5 Billeter,M., Qian,Y.Q., Otting,G., Müller,M., Gehring,W. and Wüthrich,K. (1993) *J. Mol. Biol.*, **234**, 1084–1093.
6 Hirsch,J.A. and Aggarwal,A.K. (1995) *EMBO J.*, **14**, 6280–6291.
7 Li,T., Stark,M.R., Johnson,A.D. and Wolberger,C. (1995) *Science*, **270**, 262–269.
8 Bürglin,T. (1994) *Guidebook to the Homeobox Genes*. Oxford University Press, Oxford, UK.
9 Bairoch,A. and Apweiler,R. (1998) *Nucleic Acids Res.*, **26**, 38–42.
10 Baxevanis,A.D. (1998) In *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. John Wiley & Sons, New York, NY.
11 Qian,Y., Billeter,M., Otting,G., Müller,M., Gehring,W. and Wüthrich,K. (1989) *Cell*, **59**, 573–580.
12 Billeter,M., Qian,Y., Otting,G., Müller,M., Gehring,W.J. and Wüthrich,K. (1990) *J. Mol. Biol.*, **214**, 183–197.
13 Guntert,P., Qian,Y.Q., Otting,G., Müller,M., Gehring,W. and Wüthrich,K. (1991) *J. Mol. Biol.*, **217**, 531–540.
14 Otting,G., Qian,Y.Q., Billeter,M., Müller,M., Affolter,M., Gehring,W.J. and Wüthrich,K. (1990) *EMBO J.*, **9**, 3085–3092.
15 Liu,B., Kissinger,C. and Pabo,C. (1990) *Biochem. Biophys. Res. Commun.*, **171**, 257–259.
16 Tucker-Kellogg,L., Rould,M., Chambers,K., Ades,S., Sauer,R. and Pabo,C. (1997) *Structure*, **5**, 1047–1054.
17 Hogue,C.W.V. and Bryant,S.H. (1998) In *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. John Wiley & Sons, New York, NY.
18 Hogue,C., Ohkawa,H. and Bryant,S. (1996) *Trends Biochem. Sci.*, **21**, 226–229.