The Homeodomain Resource: sequences, structures, DNA binding sites and genomic information

Sharmila Banerjee-Basu, Daniel W. Sink and Andreas D. Baxevanis*

Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA

Received October 3, 2000; Accepted October 25, 2000

ABSTRACT

The Homeodomain Resource is an annotated collection of non-redundant protein sequences, three-dimensional structures and genomic information for the homeodomain protein family. Release 3.0 contains 795 full-length homeodomain-containing sequences, 32 experimentally-derived structures and 143 homeobox loci implicated in human genetic disorders. Entries are fully hyperlinked to facilitate easy retrieval of the original records from source databases. A simple search engine with a graphical user interface is provided to query the component databases and assemble customized data sets. A new feature for this release is the addition of DNA recognition sites for all human homeodomain proteins described in the literature. The Homeodomain Resource is freely available through the World Wide Web at http://genome.nhgri.nih.gov/homeodomain.

INTRODUCTION

The homeodomain is a common DNA-binding structural motif found in many eukaryotic regulatory proteins (1,2). Homeodomain proteins are involved in the transcriptional control of many developmentally important genes, and 143 human loci have been linked to various genetic and genomic disorders. X-ray crystallographic and NMR spectroscopic studies (3–11) on several members of this family have revealed that the homeodomain motif is comprised of three α-helices that are folded into a compact globular structure. Helices I and II lie parallel to each other and across from the third helix. This third helix is also referred to as the 'recognition helix', as it confers the DNA-binding specificity of individual homeodomain proteins. The homeodomain has been evolutionarily conserved at the structural level (12); this is most evident upon examination of divergent members of the homeodomain family.

The Homeodomain Resource represents a comprehensive collection of information about the homeodomain family. The database contains all available full-length and homeodomain-only sequence data and structures as of October, 2000. The genetic data contained in this database includes information on human diseases in which homeodomain-containing proteins are implicated, cytogenetic map locations and specific mutation data underlying the disease condition. Since its last release (13), 27 new loci for genetic disorders have been identified. The sequence information within the database is automatically updated on a monthly basis, with each entry in the database rigorously selected to assure non-redundancy.

DATABASE DESCRIPTION

The current version of the database contains 795 full-length homeodomain protein sequences isolated from 83 different species (Table 1). The complete full-length sequence data as well as the homeodomain portion of the sequence is available in FASTA format. The database can be searched on the basis of SWISS-PROT ID, GenBank accession number, gene names (both common and alternative), protein description, sequence and organism name. The search engine also supports Boolean queries, allowing users to search on individual fields or on all fields at once. Search results are returned in a tabular format, with hyperlinks to the original records in GenBank and SWISS-PROT, respectively. Individual sequences can also be retrieved in FASTA format from a pop-up window.

Table 1. Homeodomain Resource statistics

Total sequences available	3796
Non-redundant full-length sequences	795
Genes/gene symbols	375
Distinct organisms	83
Three-dimensional structures	32
Homeobox loci implicated in human genetic disorders	143

The genetic information available for the homeodomain protein family has increased ~23% since the last release, with the inclusion of 27 new loci. The genetic data are compiled from both the literature and from the Online Mendelian Inheritance in Man (OMIM) database at NCBI (http://www.ncbi.nlm.nih.gov/Omim/). As before, a search engine is available for querying this genomic information. Search results are presented in a tabular format with hyperlinks to the original records and can be sorted by disease name, map location, gene symbol, protein name or OMIM identifier. This value-added format allows users easy access to related information.

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

Common Name	DNA Binding Sequence	Common Name	DNA Binding Sequence
ARIX	GT**CAAT**TAG	OCT 1	AA**ATGCAAAT**ACC
BARX2	YY**TAAT**GRTTTTY	OCT 2	**ATGCAAAT**
CDX1	CT**TTAT**G	OCT3	**ATGCAAAT**
CDX2	A A CT**TTAT**G	OTX2	CC**TAAT**CCTGGG**TTATC**
CDX4	A T**TTAT**T	PAX2	**C** G **TCAT**GCATGAC
CRT1	**TAAT**NNNATTA	PAX3	**TAAT**NN**ATTA**
CRX	C **TAAT**CA	PAX4	**C** ANNN**TCAC**CC
CSX	T**NAAG**TG	PAX5	A C T G**TCAC**GCTTGATGC
DLX2	CAC**TAAT**TGAG	PAX6	A C TA T AANNT**TCACGC**TTGATGCAC
DLX3	CAN**TTAT**CTCAG	PAX7	**TAATT**
DLX4	TCAAT**TAAT**TGA	PBX1	TTGAT**TGAT**
DLX5	A**TAAT**TAG	PBX1-HOXA10	A**TGATTTAT**GA
DLX6	A**TAAT**TAG	PBX1-HOXA9	**T** **TGATTTAC**
EMX2	CAC**TAAT**TGAG	PBX1-HOXB1	A**TGATTGAT**CG
EVX1	TN**ATTA**NNNNNTAATNG	PBX1-HOXB4	A**TGATTGAT**GA
GBX2	AGTGAG**TAAT**TGG	PBX1-HOXB6	A**TGATTTAT**TA
HEX	**T** **T** **CAAG**	PBX1-HOXB7	A**TGATTTAT**GCTCTA
HK31	**TAAG**TA	PBX1-HOXB8	A**TGATTTAT**GACTCTA
HLX1	TCTAT**TAAT**TGA	PBX1-HOXB9	A**TGATTTAC**GAC
HME1	TCTG**TAAT**TACA	PBX1-HOXC6	G A A**TGATTTAT**TGCTTT
HNF1	AGT**TTAT**ATTTGACA	PBX1-HOXC6	A G**TGATTTAT**TACTTT
HNF4	CGCTTGGCAAAGGTCACCT	PBX1-MEIS1	G **TGATTGACAG**CT
HNF6	GA**TATT**GATTTTT	PBX2	TTGAT**TGAT**
HOX 11	CGT**TAAT**TGG	PBX2-HOXC6	G**TGATTTATT**GCTTT
HOX11L1	CGT**TAAT**TGG	PBX3	TTGAT**TGAT**
HOX11L2	CGT**TAAT**TGG	PBX3-HOXC6	G G**TGATTTATT**ACTTT
HOXA10	**T** **TATT**GA	PITX1	**TAAT**CC
HOXA9	**T** **TTTC**	PITX2	**TAAT**CC
HOXB7	T**TTAT**GAC	PMX1	**TAAT**A
HOXB8	T**TTAT**GAC	PNX1	**TGACAG**
HOXB9	**T** T**TAAC**GAC	POU1F1	**TAAAT**
HOXC11	T **TTTA**C	POU2F1	**ATGC**AAAT
HOXD12	**T** T**TAAC**GAC	POU3F1	TGGA**GANAATAGAG**
HOXD13	T**TTAC**GAG	POU3F2	**ATGC**AAAT
IPF1	T CTC**TAAT**GGG	POU3F4	**ATGC**AAAT
ISL-1	GCT**TAAT**ATCTG	POU4F1	A T GCTC**ATTA**AC
ISL-2	**T**GG T**TAAG**TA	POU4F2	A T GCTC**ATTA**AC
LH2	T**TAAT**TA	POU4F3	A T GCTC**ATTA**AC
MEI1	G **TGACAG**CT	POU5F1	**ATGC**AAAT
MOX2	TCAAT**TAAT**TGA	POU6F1	**TAAT**GAGCTGCX**TAAT**
MSX-1	C**TAAT**TG	SATB1	C A**TAAT**A
MSX-2	TCAAT**TAAT**TGA	TCF-1	AACAAAG
N-OCT5	**ATGC**AAAT	TGIF	**TGACAA**
NK61	T**TAAT**TAC	TTF-1	T**CAAG**TGTT
NKX2.2	**T** A C T**CAAG**TGGTT		

Figure 1. DNA binding sites for human homeodomain loci implicated in genetic or genomic disorders. Data is derived from the published literature, citations in OMIM, and entries for DNA-bound homeodomain structures from the Protein Data Bank. The core regions of each of the DNA binding sites are shown in bold type. An expanded version of this table is available online (at http://genome.nhgri.nih.gov/homeodomain); the expanded version includes a listing of alternative gene names, as well as citation information.

A new feature of the Homeodomain Resource is the inclusion of DNA binding sites for human homeodomain loci implicated in genetic or genomic disorders (Fig. 1). These data were obtained by extensive review of the published literature, citations in OMIM, and entries for DNA-bound homeodomain structures from the Protein Data Bank. The online version of Figure 1 (available via http://genome.nhgri.nih.gov/homeodomain) includes alternate gene names and references to the primary citation from which the information was retrieved.

REFERENCES

1. Gehring,W., Qian,Y., Billeter,M., Furukubo-Tokunaga,K., Schier,A., Resendez-Perez,D., Affolter,M., Otting,G. and Wuthrich,K. (1994) Homeodomain-DNA recognition. *Cell*, **78**, 211–223.

2. Laughon,A. (1991) DNA binding specificity of homeodomains. *Biochemistry*, **30**, 11357–11367.
3. Dekker,N., Cox,M., Boelens,R., Verrijzer,C., van der Vliet,P. and Kaptein,R. (1993) Solution structure of the POU-specific DNA-binding domain of Oct-1. *Nature*, **362**, 852–855.
4. Viglino,P., Fogolari,F., Formisano,S., Bortolotti,N., Damante,G., Di Lauro,R. and Esposito,G. (1993) Structural study of rat thyroid transcription factor 1 homeodomain (TTF-1HD) by nuclear magnetic resonance. *FEBS Lett.*, **336**, 397–402.
5. Gruschus,J., Tsao,D., Wang,L., Nirenberg,M. and Ferretti,J. (1997) Interactions of the vnd/NK-2 homeodomain with DNA by nuclear magnetic resonance spectroscopy: basis of binding specificity. *Biochemistry*, **36**, 5372–5380.
6. Kissinger,C., Liu,B., Martin-Blanco,E., Kornberg,T. and Pabo,C. (1990) Crystal structure of an engrailed homeodomain-DNA complex at 2.8 Å resolution: a framework for understanding homeodomain-DNA interactions. *Cell*, **63**, 579-590.
7. Ceska,T., Lamers,M., Monaci,P., Nicosia,A., Cortese,R. and Suck,D. (1993) The X-ray structure of an atypical homeodomain present in the rat liver transcription factor LFB1/HNF1 and implications for DNA binding. *EMBO J.*, **12**, 1805–1810.
8. Liu,B., Kissinger,C. and Pabo,C. (1990) Crystallization and preliminary X-ray diffraction studies of the engrailed homeodomain and of an engrailed homeodomain/DNA complex. *Biochem. Biophys. Res. Commun.*, **171**, 257–259.
9. Qian,Y., Billeter,M., Otting,G., Muller,M., Gehring,W. and Wuthrich,K. (1989) The structure of the Antennapedia homeodomain determined by NMR spectroscopy in solution: comparison with prokaryotic repressors. *Cell*, **59**, 573–580.
10. Qian,Y., Furukubo-Tokunaga,K., Resendez-Perez,D., Muller,M., Gehring,W. and Wuthrich,K. (1994) Nuclear magnetic resonance solution structure of the fushi tarazu homeodomain from Drosophila and comparison with the Antennapedia homeodomain. *J. Mol. Biol.*, **238**, 333–345.
11. Wolberger,C., Vershon,A., Liu,B., Johnson,A. and Pabo,C. (1991) Crystal structure of a MAT alpha 2 homeodomain-operator complex suggests a general model for homeodomain-DNA interactions. *Cell*, **67**, 517–528.
12. Buerglin,T. (1994) In Duboule,D. (ed.), *Guidebook to the Homeobox Genes*. Oxford University Press, Oxford, UK, pp. 25–71.
13. Banerjee-Basu,S., Ryan,J.F. and Baxevanis,A. (2000) The Homeodomain Resource: sequences, structures and genomic information. *Nucleic Acids Res.*, **28**, 329–330.