Modern Internet Technologies for Hypermedia Information Systems Describing Ensembles of Genes-controllers of Development

Alexander V. Spirov
The Sechenov Institute of Evolutionary Physiology and Biochemistry,
Russian Academy of Sciences, St. Petersburg, Russia
Spirov@iephb.ru

Abstract

Nowadays homeobox genes were found for species of main invertebrate and vertebrate phyla. It is established that these genes play key role in time and space orchestration of genome expression during development. Auto- and crossregulatory functional interactions join homeobox genes into genetic networks. For the purpose to orange all available data on structure, functions, phylogeny and evolution of *Hox*-genes, *HOX*-clusters and *Hox*-networks we develop specialized database *HOX*-Pro. Its main location is http://www.iephb.ru/~spirov/hox_pro/hox-pro00.html.
The DB is also mirrored at
http://www.mssm.edu/molbio/hoxpro/new/hox-pro00.html.

HOX Pro has been designed not as an electronic table but as an interactive hypermedia information system embedded in the Internet. Such Web resources begin at major search engines, include the database servers and finish at the end-user's computer. The HOX Pro graphical user interface, written in Java, allows exploration and visualization of the database through the Internet. It includes tools for automated generation of the gene network diagrams, visualization filters, as well as tools for data navigation.

These include interactive images in the diagram, online help, interactive cross references within this database, and references to other databases. To make HOX Pro as easy to use as possible, we use javascript to orient the user by means of automated pop-up windows. These small windows automatically open when a user connects with particular html-page and contain, depending on the page, either brief DB structure map, or map of genetic ensemble. Such small help-pages assist user who are novices to the DB structure.

The distinctive feature of the HOX Pro is its ability to serve not only as an information resource but also as tool for derivation of new knowledge by means of computational analysis. The long-range goal of HOX Pro is the functional reconstruction and prediction of genetic regulatory pathways from genomic sequence information.

1. Introduction

Genome sequencing projects for a variety of organisms have resulted in rapidly enlarging catalogs of genes and gene products. The next obvious step is to establish the functional implications of these data; that is to establish both experimentally and computationally when, where, and how genes and molecules function in living organisms. In fact, our knowledge on the functioning of genes and molecules is also rapidly expanding owing to the advancement of experimental technologies in wide areas of molecular and cellular biology. In order to make full use of the information obtained by the genome projects, it is essential that such functional data are properly processed, stored and retrieved.

The functional data that relate to sequence information are currently stored mainly in the feature tables of the sequence databases and in the motif libraries [1-2] relationships of single molecules/genes and they do not contain higher level information of genetic and molecular interactions. The investigation of regulatory regions of eukaryotic genes involved in transcription control requires the creation of special databases containing exhaustive information about the structural-functional organization of these regions in terms of their interaction with the protein products of genes involved in transcription regulation [3-4]. A further complication for developmentally important genes is that function must be represented at multiple levels ranging from the molecular to the organismal.

In recent years it has been demonstrated that the protein products of a relatively small group of genes is of particular

Proceedings of the Workshop on Computer Science and Information Technologies CSIT'99
Moscow, Russia, 1999

importance in controlling the expression of a much wider set of target genes, and through them the overall course of development and morphogeneisis. An important subset of these genes contain a 180 bp structural motif known as the "homeobox", and among these an important subset occur in conserved clusters ("HOX complexes") on the chromosome [5]. Members of HOX complexes are of particular importance in specifying the overall animal body plan, and have been the object of intensive study. For these reasons, the homeobox containing genes are a natural choice for the subject matter of a database concerned with gene function in development at multiple levels.

The project which we describe here, the "Homeobox Gene Promoter Regions DataBase" (HOX Pro DB), is aimed at integrated computer presentation of current knowledge of molecular aspects of modern developmental biology in terms of the information pathways from genes to developing organs and tissues.

HOX Pro is a hypermedia information repository accessed over the World Wide Web, but it is not merely a static resource to be searched and browsed. The distinctive feature of the HOX Pro is its ability to serve not only as an information resource but also as tool for derivation of new knowledge by means of computational analysis. The long-range goal of HOX Pro is the functional reconstruction and prediction of genetic regulatory pathways from genomic sequence information.

2. General description of HOX Pro

We developed the HOX Pro database for the description of ensembles of homeobox-genes which control embryogenesis [6]. It contains a broad spectrum of information including pictures, schemes, and movies. Graphical representation of HOX clusters and HOX based networks is accomplished by way of flow diagrams, JavaScript animation, and Java applets. This permits the clear representation of gene interactions in the HOX gene ensembles and facilitates navigation in the database [7].

The HOX Pro database contains data on the structural and functional organization of the transcription regulatory machinery of homeobox and functionally related genes. The hierarchical organization of transcription regulation of metazoan genes is put into the database schema. HOX Pro also includes a hypertextual description of mechanisms of homeobox-genes activation as well as the functional characteristics of proteins encoded by homeobox-containing and functionally related genes. HOX Pro also contains links on others databases such as TRANSFAC, COMPEL, EPD, EMBL, GeNet, FlyBase, and Interactive Fly.

The main principles of data presentation in HOX Pro are as follows. The genetic cluster and genetic network maps form the basis for information structuring in HOX Pro. The genetic network diagrams are represented in a form of an oriented graph, in which each gene is represented as a node. Each node is represented by a symbol denoting its broad functional class (homeobox containing or not, controller or target, etc).

HOX Pro is subdivided into sections, which hold information on HOX clusters and networks in different organisms: *C. elegans*, the sea urchin *S. purpuratus*, the fuit fly *D. melanogaster*, chicken, mouse, human, and several other vertebrates. Each section contains up to 6 types of data: genetic cluster maps, network maps, gene entries, gene sequence entries, regulatory region entries and bibliography.

Each gene entry contains information on the gene's function, key features of its encoded protein, expression pattern, regulatory interactions (upstream and downstream genes) as well as links to other databases. The regulatory element entry in HOX Pro contains data on the organism source, bibliographical data, regulatory element sequence and coordinates of sites for transcription factors binding, as well as key words and definition. Gene interaction entries hold information on the mechanism of gene interaction with experimental lines of evidence supporting the named mechanism.

The HOX Pro graphical user interface, written in Java, allows exploration and visualization of the HOX Pro database through the Internet. It includes tools for automated generation of the gene network diagrams, visualization filters, as well as tools for data navigation. These include interactive images in the diagram, online help, interactive cross references within this database, and references to other databases [7].

One of the difficulties in the design and use of a database of this type is that for some genes there are detailed data on the structure of their regulatory regions, including binding sites for transcription factors, while for others there are other types of experimental data which may be more qualitative. Such data might include the analysis of mutations in gene regulatory regions, functional relationships with other genes, etc. A flexible data representation scheme in HOX Pro allows these difficulties to be circumvented. In accordance with the available experimental data, the structure of the transcriptional regulatory regions of some genes is given in a highly structured format, while regulation of the expression of others is described in a freeform manner with hypertext, figures and tables.

Three entrance html-pages are provided allowing the user to browse the database, to search the database, and to work with genetic cluster or network maps. While browsing HOX Pro the user sequentially moves from the page containing the list of database sections to genetic cluster and network maps. Each gene of such map is linked to gene entry, which in turn holds hypertext links to data on gene sequence, regulatory regions, gene interactions and bibliography. Thus by clicking on gene name in the map the user gets detailed information about gene and mechanisms of its regulation.

Another entrance page into HOX Pro enables the user to work with genetic cluster or network maps. Cluster

interactive diagrams present a physical map of a HOX cluster. Genetic network maps depicted as diagrams enable the user to find out which genes regulate a given gene as well as or which genes are its regulatory targets. Each gene on the cluster or network diagram is hyperlinked to its gene entry so that the user can retrieve all the information about gene of interest following the hypertext links in the database.

From the outset, HOX Pro has been designed not as an electronic table but as an interactive hypermedia information system embedded in the Internet. Such Web resources begin at major search engines, include the HOX Pro servers and finish at the end-user's computer. Four years of HOX Pro server statistics show that most users find its html pages via search engines. Hence the effective presentation of all HOX Pro resources at the most popular search engines is important. Hence we allow registration of all our html pages in the search engines, carefully control the resume of the main HOX Pro pages at the search engines, and include lists of keywords in the HOX Pro pages.

To make HOX Pro as easy to use as possible, we use javascript to orient the user by means of automated pop-up windows. These small windows automatically open when a user connects with particular html-page and contain, depending on the page, either brief DB structure map, or map of HOX/HOM cluster or network. Such small help-pages assist user who are novices to the DB structure. In addition, all 200 gene title pages include a thumbnail image of the diagram of the network and/or cluster which contains the given gene. The thumbnail is linked to the clickable diagram of the network or cluster.

The current HOX Pro version includes approximately 600 html-pages plus over 300 images occupying 7 Mb of disk space. It contains information on 200 genes and 90 promoters which are linked to maps of 13 HOX clusters and 9 genetic networks.

3. Discussion and Conclusion

The large-scale projects to sequence the DNA of humans and several other organisms has led to a rapid growth of biological information [8-10], much of it accessed in databases. A great many of these databases are designed for the storage, processing and retrieval of molecular biology data. At present neither analysis of results, nor planning of experiments are impossible without handling of these databases. Now, at the beginning of "post-genome era", when biomedical researches shift from identifying genes to characterization of their function, the design of databases containing functional information becomes crucial.

As biomedical research shifts from the identification of genes to the characterization of their function, the design of databases containing functional information becomes crucial. In a manner similar with other functional databases HOX Pro contains functional information on mechanisms of genes action in embryogenesis. However the distinctive feature of HOX Pro is a model for information presentation, which is based on a concept of genetic ensembles (clusters and networks) and a comparative evolutionary approach. Such a database structure enables end users to retrieve information on the functional organization and evolutionary conservation of a whole ensemble of interacting genes. Another distinctive feature of the HOX Pro is its user-friendly decentralized architecture, so that from any particular html-page user can see the main content of the base.

Nowadays the *Hox* genes have become a paradigm for the conservation of developmental mechanisms throughout the animal kingdom. They encode transcription factors that act as molecular markers for the position of cells along the major body axis. Individual *Hox* genes are activated at different positions in the early embryo, establishing a pattern that is maintained throughout much of development. This differential expression has been shown to control the development of region-specific structures in nematodes, arthropods and chordates, and may be a shared characteristic of triploblastic metazoan animals.

Mutations within homeotic *Hox*-genes in *Drosophila melanogaster* transform defined segments of the body into the character of adjacent segments. Homologous genes there are in vertebrates, where it has been possible to mutate them, shifts in morphological character have also been produced.

The discovery of the first *HOX* cluster outside *Drosophila* causes great hopes. It was proposed that the homeobox would become a "Rosetta stone" for the study of animal development. This could enable us to read the epigenetic code of other animals on the basis of our understanding of *Drosophila* [11]. It is possible that this *HOX*-based epigenetic code is very ancient and was in place in the common ancestor of all modern animals.

The *Hox* gene family is highly conserved yet is responsible for the development of many novel features of the vertebrate body plan. The *Hox* gene family shows progressive expansion by gene duplication in invertebrate species as a tightly linked gene cluster, and furthers amplification within the chordates primarily by cluster duplication.

As a first step towards a developmental history of animal architectures, we begin to reconstruct the evolution of the *HOX* clusters using information from developmental/molecular. Essential feature of HOX ensembles is their conservative cluster organization, suggesting that the clustering may have functional significance. In other words, functional organization of the hox-genetic networks is under control of physical structure of the chromosomal HOX cluster. In this regards, the data accumulated in the HOX Pro give us possibility to address such questions as [12]:

Do common mechanisms of transcriptional regulation exist among hox genes?

Are *cis*-elements organized into particular patterns?

Can unique sequence features be identified?

The massive of data collected in the HOX Pro allows giving positive answers on all three these questions.

Acknowledgments

This work is supported by Russian Foundation for Basic Researches (Grants No 98-04-49422 and No 98-07-90373).

References

1. Bucher N. "EPD: Eukaryotic promoter database". Current release. 1989.
2. Ghosh D. "Status of the transcription factors database (TFD)". *Nucl Acids Res* 1993; 21:3117-3118.
3. Wingender E, Kel A.E, Kel O.V, Karas H, Heinemeyer T, Dietze P, Knuppel R, Romaschenko A.G, Kolchanov N.A. "TRANSFAC, TRRD and COMPEL: towards a federated database system on transcriptional regulation". *Nucleic Acids Res* 1997; 25:265-8.
4. Kanehisa, M. "Toward pathway engineering: a new database of genetic and molecular pathways". *Science & Technology Japan* 1996; 59:34-38.
5. McGinnis W., Krumlauf R. "Homeobox genes and axial patterning". *Cell* 1992; 68:283-302.
6. Spirov A.V. "The role of some conservative sequences in regulatory elements of antp-like, homeobox-containing genes of vertebrates". *J Evol Biochem and Physiol* 1996; **32**:556-564.
7. Serov V.N., Spirov A.V. and Samsonova M.G. "Graphical interface to genetic network database GeNet". *Bioinformatics* 1998; 14:546-547.
8. Lander E.S. "The new genomics: Global view of biology". *Science* 1996; 274:536.
9. Nowak R. "Entering in postgenome era". *Science* 1995; 270:368-369.
10. Schuler G.D, Boguski M.S., Stewart L.D. et al. "A gene map of the human genome". *Science* 1996; 274:540.
11. Slack J.M., P.W. Holland and C.F. Graham "The zootype and the phylotypic stage" *Nature* 1993; **361**:490.
12. Kappen C., Schughart K., Ruddle F.H. "Two steps in the evolution of Antennapedia-class vertebrate homeobox genes". *Proc Nat Acad Sci USA.* 1989; 86:5459-5463.