GIF-DB, a WWW database on gene interactions involved in *Drosophila melanogaster* development

Bernard Jacq*, Florence Horn, Florence Janody, Nicolas Gompel, Olivier Serralbo, Elodie Mohr, Christine Leroy, Bernard Bellon[1], Laurent Fasano, Patrick Laurenti[2] and Laurence Röder

Laboratoire de Génétique et Physiologie du Développement, IBDM, Parc Scientifique de Luminy, CNRS Case 907, 13288 Marseille Cedex 09, France, [1]Atelier de Bio-Informatique, Case 13, Université de Provence, 3 Place Victor Hugo, 13331 Marseille Cedex 03, France and [2]Laboratoire de Biologie du Développement (Anatomie Comparée), Université de Paris VII, case 7077, 75251 Paris Cedex 05, France

Received September 23, 1996; Accepted September 24, 1996

ABSTRACT

GIF-DB (Gene Interactions in the Fly Database) is a new WWW database (http://www-biol.univ-mrs.fr/~lgpd/GIFTS_home_page.html) describing gene molecular interactions involved in the process of embryonic pattern formation in the fly *Drosophila melanogaster*. The detailed information is distributed in specific lines arranged into an EMBL- (or SWISS-PROT-) like format. GIF-DB achieves a high level of integration with other databases such as FlyBase, EMBL and SWISS-PROT through numerous hyperlinks. The original concept of interaction databases examplified by GIF-DB could be extended to other biological subjects and organisms so as to study gene regulatory networks in an evolutionary perspective.

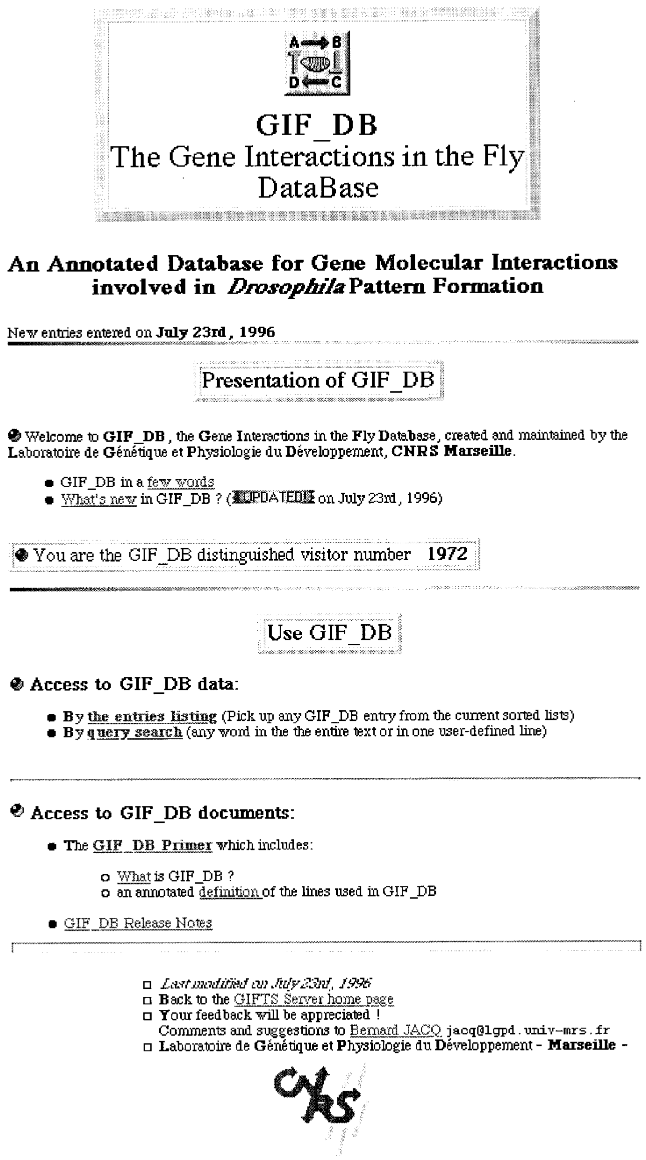
INTRODUCTION

Databases are now of widespread use in biological research and this issue of *Nucleic Acids Research* provides the reader with an up-to-date collection of different biological databases which have various scientific purposes and contents. The majority of these databases can be classified as mainly structural in that the core of their informational content is based on various aspects of DNA, RNA or protein sequence and/or structure. Relatively few databases have a content and an organization which are oriented towards the biological function of the genes and the relationships between structure and function. EcoCyc, an encyclopedia of *Escherichia coli* genes and metabolism (1) is an example of a database which integrates functional aspects as one can find both data on gene structure and their function in the regulation of biochemical pathways. In the field of genetic diseases, OMIM, a catalog of human genes and genetic disorders (2) provides the user with both structural (molecular genetics, biochemistry, genetic mapping) and functional data (clinical features, diagnosis, inheritance, etc.) on human genes.

We are interested in the biological process of pattern formation in *Drosophila* and in understanding the basis of specific identity acquisition by the different body parts (3–7). In *Drosophila*, different classes of genes involved in the segmentation processes (maternal, gap, pair-rule and segment polarity genes) divide the embryo along the antero-posterior axis into repeated homologous units (8,9), which will develop specific identities and morphogenetic features under the control of homeotic genes (10). Specific interactions within and between these gene families are essential for the establishment of a correct body pattern. Being able to access, query and manipulate the data on these developmental genes and their functional interactions within specific regulatory networks is now an important need for developmental and molecular biologists studying gene regulation.

Gene molecular interactions, i.e direct molecular interactions involving DNA, RNA and proteins, play an essential part in all known biological processes. Although different databases exist for each of these three types of macromolecules, data concerning precise molecular interactions between them are underrepresented in these databases. If one considers protein/DNA interactions for instance, only four examples of such co-crystals are found (of 15 homeodomain structures) in PDB, an archive of experimentally determined three-dimensional structures of biological macromolecules (11). In addition, it is extremely difficult to extract from GenBank (12), EMBL (13), PIR-international (14) or SWISS-PROT (15) databases a list of proteins which interact with a given gene or a list of target genes for a given DNA-binding protein and the same is true for protein/RNA and protein–protein interactions. The Transfac database (16) gives some precise structural data for transcription factors and their known binding sites. Even in this case, however, data essential for the understanding of transcription factor function in their specific biological contexts are missing: developmental stage at which interaction occurs, phenotype of animals in which the transcription factor is absent or mutated, biological result of the interaction, organisation of the *cis*-regulatory region, experimental evidence for interaction.

*To whom correspondence should be addressed. Tel: +33 491 26 90 55; Fax: +33 491 82 06 82; Email: jacq@lgpd.univ-mrs.fr



GIF_DB
The Gene Interactions in the Fly
DataBase

An Annotated Database for Gene Molecular Interactions involved in *Drosophila* Pattern Formation

New entries entered on **July 23rd, 1996**

Presentation of GIF_DB

● Welcome to **GIF_DB**, the **G**ene **I**nteractions in the **F**ly **D**atabase, created and maintained by the **L**aboratoire de **G**énétique et **P**hysiologie du **D**éveloppement, **CNRS Marseille**.

- GIF_DB in a few words
- What's new in GIF_DB ? (UPDATED!! on July 23rd, 1996)

● You are the GIF_DB distinguished visitor number **1972**

Use GIF_DB

● **Access to GIF_DB data:**

- **By the entries listing** (Pick up any GIF_DB entry from the current sorted lists)
- **By query search** (any word in the the entire text or in one user-defined line)

● **Access to GIF_DB documents:**

- The **GIF_DB Primer** which includes:
  - ○ What is GIF_DB ?
  - ○ an annotated definition of the lines used in GIF_DB
- GIF_DB Release Notes

□ *Last modified on July 23rd, 1996*
□ **B**ack to the GIFTS Server home page
□ **Y**our feedback will be appreciated !
Comments and suggestions to Bernard JACQ jacq@lgpd.univ-mrs.fr
□ **L**aboratoire de **G**énétique et **P**hysiologie du **D**éveloppement - **Marseille** -

Figure 1. The GIF-DB home page.

In this paper, we describe the concepts, organization, content and use of GIF-DB, the Gene Interactions in the Fly Database, a new WWW database which aims at providing a repository for data on gene interactions involved in *Drosophila* embryonic development and the regulatory networks in which they are implied.

LEADING CONCEPTS OF GIF-DB

Four main leading concepts were considered to elaborate GIF-DB.

Detailed and structured description of interactions

Our aim was to find a relatively simple, but well defined way to represent the various and complex knowledge we presently have on gene molecular interactions during embryonic development of *Drosophila*. This led to the conception of a structured entry format which is described in the next chapter.

Integration of GIF-DB data with that of other WWW databases

As GIF-DB was designed to be accessible on the web, it was of great importance that it essentially includes original data and relies on other databases to access related data already described elsewhere. This goal is achieved through hyperlinks pointing towards external molecular and genetic databases. At the moment, hyperlinks towards three different databases have been introduced: EMBL, SWISS-PROT and FlyBase, the genetic and molecular *Drosophila* database (17). In this latter case, links are pointing either towards the gene entry (FBgn in FlyBase) or the bibliographic reference (FBrf in FlyBase) (Fig. 1).

Classification of all interactions in one of three major interaction types

Gene molecular interactions should not be mistaken for genetic interactions. The latter ones are more general and include both indirect and direct interactions. Our working definition is: there is a direct molecular interaction between gene A and gene B if gene A or one of its products (i.e. mRNA or protein) physically interacts at the molecular level with gene B or one of its products (mRNA or protein). In GIF-DB, we have focused on direct gene interactions, and six different molecular types of interaction could theoretically be considered: DNA–DNA, DNA–RNA, RNA–RNA, DNA–protein, RNA–protein and protein–protein interactions. Among these possibilities, we will consider further three major types of interactions only, which are by far the most documented ones, whatever the organism being considered: (i) protein–DNA interactions (type I); protein–RNA interactions (type II); and protein–protein interactions (type III).

A practical consequence of the above definition of interaction types is that we will only take into account binary interactions (i.e. interactions occurring between two molecular partners). This could be viewed as a limitation if one considers what is already known about the complexity of gene interactions. However, and within certain limits, any complex interaction which involves more than two partners (interaction between a DNA sequence and several proteins, or between several proteins into a multimeric complex, for instance) could be split up into several binary interactions in order to be described.

Generic mode of interaction representation

Although our purpose in GIF-DB is to focus on interactions involved in the biological process of *Drosophila* pattern formation, we designed the file structure of GIF-DB so that it could have a generic value. We therefore propose herein a multipurpose tool for the representation of interaction knowledge. Our aim was to create a general format which could be used for the description of nearly any gene interaction, whatever the biological process and the organism in which they occur may be.

DATABASE ORGANIZATION

The GIF-DB interaction database is a collection of hypertext files, each of them describing an interaction between two partners as discussed above. Each entry contains biological information which has been arranged into an 'EMBL-like' or 'SWISS-PROT-like' model format. Several reasons have dictated such a choice. (i) The EMBL and SWISS-PROT formats, which are quite similar, meet simplicity, logic and power of data representation. (ii) The complementarity between the description of nucleic acid and protein data found in these two European databases is a concept that is useful for a new database in which protein, DNA and RNA data will be found altogether. (iii) Finally, adhering closely to an already existing data structure model will allow future users to find themselves into a relatively familiar environment, even if some obligatory differences will exist due to the different nature of our database.

As is the case for EMBL and SWISS-PROT entries, GIF-DB ones are structured so as to be usable by human readers as well as by computer programs. The explanations, descriptions, classifications and other comments are in ordinary English. Wherever possible, symbols and nomenclature supposed to be familiar to drosophilists, geneticists, biochemists and molecular biologists are used to describe the interactions and some conventions used in FlyBase have been followed.

All data found in GIF-DB comes from the literature. The information coming from different papers is compiled (and synthetized if necessary), verified and entered in DEXIFLY, a relational *Drosophila* database (Horn *et al.*, in preparation). The HTML files constituting GIF-DB are then automatically generated from this database.

Finally, a user manual (The GIF-DB primer) is accessible on-line, explaining what is GIF-DB, with a particular emphasis on the problem of interaction data representation and describing in detail all the line types and adopted conventions.

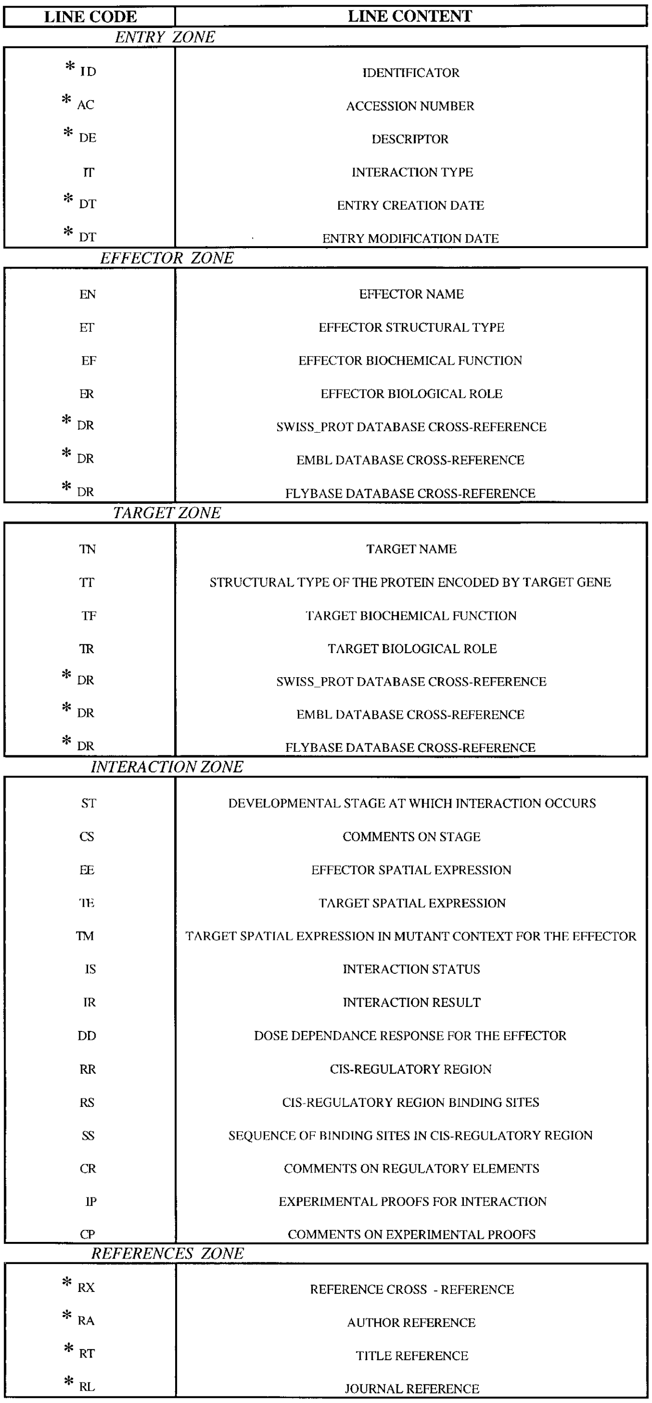
ENTRY FORMAT

Each entry in the GIF-DB database is composed of lines and at the moment, only entries describing protein/DNA interactions (type I) have been entered in the database. Different types of lines (each having its own format) are used to record the various types of gene interaction information which make up the entry. As is the case for the EMBL and SWISS-PROT databases, each line in a GIF-DB entry begins with a two-character line code which indicates the type of information contained in the line. Wherever it was possible, we have tried to use the linetypes already established by the EMBL and SWISS-PROT databases.

There are two main differences between the GIF-DB format and the EMBL and SWISS-PROT ones (i) Owing to the specificity of the GIF-DB database, many new line types had to be introduced. A typical SWISS-PROT entry contains a maximum of 17 different line types whereas GIF-DB contains at the moment 40 different linetypes. (ii) Because of this important number, the different line types in a GIF-DB entry have been grouped into five zones, in order to improve human readibility. These are the ENTRY zone, the EFFECTOR zone, the TARGET zone, the INTERACTION zone and the REFERENCES zone, respectively, always in that order, each zone containing specific line types.

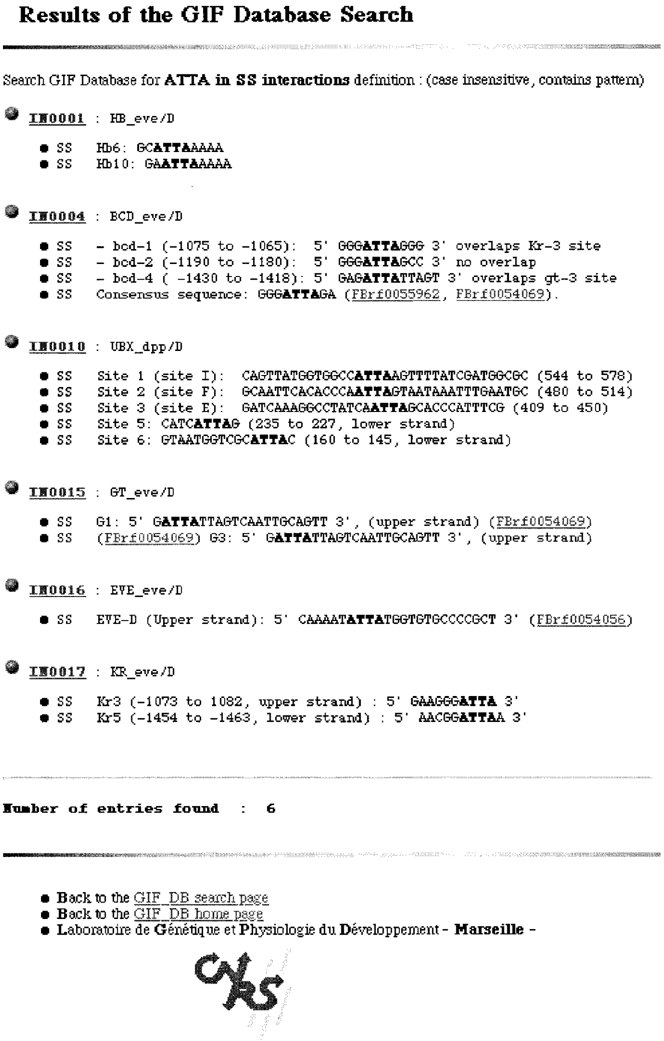
Each entry in GIF-DB is four to six text pages long so that displaying an example entry in this article was not possible. Rather, we present in Table 1 the list of all different data lines and their purpose. With the exception of the text-free lines (comment lines and the RR, RS and SS lines associated with *cis*-regulatory sequences description), all other line types have a controlled vocabulary content, i.e. only words belonging to pre-defined lists are accepted for the data description. Most of the lines are either self-explanatory or also found in EMBL or SWISS-PROT databases and therefore will not be discussed further. Some of the original line types deserve a few comments. For example, the IP line (experimental proofs for interaction) and its associated CP (comments on experimental proofs) line list all experimental

Table 1. The different line types used in GIF-DB



LINE CODE	LINE CONTENT
ENTRY ZONE	
* ID	IDENTIFICATOR
* AC	ACCESSION NUMBER
* DE	DESCRIPTOR
IT	INTERACTION TYPE
* DT	ENTRY CREATION DATE
* DT	ENTRY MODIFICATION DATE
EFFECTOR ZONE	
EN	EFFECTOR NAME
ET	EFFECTOR STRUCTURAL TYPE
EF	EFFECTOR BIOCHEMICAL FUNCTION
ER	EFFECTOR BIOLOGICAL ROLE
* DR	SWISS_PROT DATABASE CROSS-REFERENCE
* DR	EMBL DATABASE CROSS-REFERENCE
* DR	FLYBASE DATABASE CROSS-REFERENCE
TARGET ZONE	
TN	TARGET NAME
TT	STRUCTURAL TYPE OF THE PROTEIN ENCODED BY TARGET GENE
TF	TARGET BIOCHEMICAL FUNCTION
TR	TARGET BIOLOGICAL ROLE
* DR	SWISS_PROT DATABASE CROSS-REFERENCE
* DR	EMBL DATABASE CROSS-REFERENCE
* DR	FLYBASE DATABASE CROSS-REFERENCE
INTERACTION ZONE	
ST	DEVELOPMENTAL STAGE AT WHICH INTERACTION OCCURS
CS	COMMENTS ON STAGE
EE	EFFECTOR SPATIAL EXPRESSION
TE	TARGET SPATIAL EXPRESSION
TM	TARGET SPATIAL EXPRESSION IN MUTANT CONTEXT FOR THE EFFECTOR
IS	INTERACTION STATUS
IR	INTERACTION RESULT
DD	DOSE DEPENDANCE RESPONSE FOR THE EFFECTOR
RR	CIS-REGULATORY REGION
RS	CIS-REGULATORY REGION BINDING SITES
SS	SEQUENCE OF BINDING SITES IN CIS-REGULATORY REGION
CR	COMMENTS ON REGULATORY ELEMENTS
IP	EXPERIMENTAL PROOFS FOR INTERACTION
CP	COMMENTS ON EXPERIMENTAL PROOFS
REFERENCES ZONE	
* RX	REFERENCE CROSS - REFERENCE
* RA	AUTHOR REFERENCE
* RT	TITLE REFERENCE
* RL	JOURNAL REFERENCE

The first column (line code) lists the two-letter codes indicating the type of data contained in the corresponding line of each entry. They are listed in the order in which they appear and are grouped according to the five zones which make up every entry. Line codes marked with an asterisk are used with the same purpose in the EMBL and SWISS-PROT databases. The second column (line content) lists the type of data corresponding to each line code.



```
Results of the GIF Database Search

Search GIF Database for ATTA in SS interactions definition : (case insensitive, contains pattern)

IN0001 : HB_eve/D

  ● SS   Hb6: GCATTAAAAA
  ● SS   Hb10: GAATTAAAAA

IN0004 : BCD_eve/D

  ● SS   - bcd-1 (-1075 to -1065):  5' GGGATTAGGG 3' overlaps Kr-3 site
  ● SS   - bcd-2 (-1190 to -1180):  5' GGGATTAGCC 3' no overlap
  ● SS   - bcd-4 ( -1430 to -1418): 5' GAGATTATTAGT 3' overlaps gt-3 site
  ● SS   Consensus sequence: GGGATTAGA (FBrf0055962, FBrf0054069).

IN0010 : UBX_dpp/D

  ● SS   Site 1 (site I):  CAGTTATGGTGGCCATTAAGTTTTATCGATGGCGC (544 to 578)
  ● SS   Site 2 (site F):  GCAATTCACACCCAATTAGTAATAAATTTGAATGC (480 to 514)
  ● SS   Site 3 (site E):  GATCAAAGGCCTATCAATTAGCACCCATTTCG (409 to 450)
  ● SS   Site 5: CATCATTAG (235 to 227, lower strand)
  ● SS   Site 6: GTAATGGTCGCATTAC (160 to 145, lower strand)

IN0015 : GT_eve/D

  ● SS   G1: 5' GATTATTAGTCAATTGCAGTT 3', (upper strand) (FBrf0054069)
  ● SS   (FBrf0054069) G3: 5' GATTATTAGTCAATTGCAGTT 3', (upper strand)

IN0016 : EVE_eve/D

  ● SS   EVE-D (Upper strand): 5' CAAAATATTATGGTGTGCCCCGCT 3' (FBrf0054056)

IN0017 : KR_eve/D

  ● SS   Kr3 (-1073 to 1082, upper strand) : 5' GAAGGGATTA 3'
  ● SS   Kr5 (-1454 to -1463, lower strand) : 5' AACGGATTAA 3'

Number of entries found  :  6

  ● Back to the GIF_DB search page
  ● Back to the GIF_DB home page
  ● Laboratoire de Génétique et Physiologie du Développement - Marseille -
```

Figure 2. An example of a GIF-DB query search. The 'ATTA' motif was searched for through the SS lines of all entries. The result is displayed on the screen as a hypertext file (hyperlinks are underlined). The accession number and name of the entries in which a match was found are shown and each matching line is displayed with the queried chain appearing in bold. The total number of entry hits is indicated at the bottom of the page.

methods used to conclude in favor of a direct interaction. The content of these two lines therefore provides a 'likelihood coefficient' for the interaction in that the more different methods have been used, the most likely the interaction. The group of RR, RS and SS lines provides complete information on the *cis*-regulatory regions, ranging from: their location in the gene (RR lines); the number of binding sites and relative affinity if known (RS lines); and finally the sequence of the different binding sites (SS lines). This group of lines can be used as many times as necessary for the description of different *cis*-regulatory regions.

INTERACTIVE ACCESS USING THE WORLD WIDE WEB

The World Wide Web is at the moment the only way to access GIF-DB data through the GIFTS (Gene Interaction in the Fly Transworld Server) WWW server in Marseille. To access the WWW, one needs a WWW browser such as Netscape Navigator^{TM} (from Netscape Communications Corp.) and a link to the Internet. The URL (Uniform Resource Locator, the addressing system used in the WWW) of the GIFTS Server is http://www-biol.univ-mrs.fr/~lgpd/GIFTS_home_page.html . At the moment, there are two different ways to access the data once the connection with the server is established. First, a hypertext list of all available entries is accessible from the GIF-DB home page. The entries are sorted in three different ways to facilitate retrieval: by alphabetical order of effector protein, by alphabetical order of target gene and by entry accession number. A mouse click on a given entry name allows the display of the corresponding complete hypertext file for the entry. Second, a powerful query search program allows to search for the occurrence of an ASCII character chain entered by the user. The search can be performed either on the entire database or in anyone of the 40 different datalines of all entries. Two additional options allow a case-sensitive or case-insensitive search on the one hand and to have the queried chain either matching exactly one word or to be a part of it, on the other hand. An example of a query search in the SS line (binding-sites in *cis*-regulatory sequences) using the partial match option is given in Figure 2. In this particular case, the search has allowed the user to find all binding sites whose sequences contain the core sequence of homeodomain binding sites (ATTA). The result is itself an hypertext page which gives access to complete entries.

FUTURE PROSPECTS

One way of looking at the GIF-DB database is to consider that such a database of interactions provides a natural way to make functional links between entries in different molecular databases. Such functional links are a useful complement to the structural links (database cross-references) already present between an EMBL (or GenBank) entry and a SWISS-PROT (or PIR-International) corresponding translational product. In an evolutionary perspective, building interaction databases for different organisms would be extremely interesting as it would provide a means to test if homologous genes are working through homologous regulatory pathways.

Within the next 2 years, we will offer new possibilities within GIF-DB through the addition of new linetypes and the adjunction of hyperlinks towards more databases such as GenBank, PIR-International, Transfac and Flyview (18), a database on expression patterns of *Drosophila* genes. We are also presently defining a format for the description of protein/protein interactions (in which a few lines will differ from that used for protein/DNA interactions) so as to be able to enter this kind of data in the database in the near future. This will be of particular importance to describe all interactions involved in signal transduction pathways.

A few hundred molecular interactions are known at present in the field of *Drosophila* pattern formation and as each GIF-DB entry contains many different data, reaching completeness in the description of pattern interactions could be only a long-term objective. In the mean time, we plan to release a second WWW database, which will essentially provide a few hyperlinks and some bibliographic references for each interaction. This database, called DIDRO (Dictionary of Interactions in *Drosophila*) will keep track of pattern formation interactions but also of interactions involved in other aspects of *Drosophila* biology (sex determination, organogenesis, nervous system formation, etc.). Finally, the data on interactions will be included in a knowledge base, presently under development (Chemla *et al.*, in preparation), within which graphical representations of interactions and regulatory networks will be automatically generated. This will represent a first step towards the simulation of some aspects of the dynamic behavior of developmental genetic regulatory networks.

CITING GIF-DB

If you use GIF-DB as a tool in your published research work, please cite this paper.

ACKNOWLEDGEMENTS

We would like to thank C. Vermenot and S. Daloiseau for help in literature analysis, B. Mahé for a prototype version of the query search program, C. Chemla and V. Pillet for helpful comments. F. Horn is a fellow of the French MESR. This work has been supported by grants from the GREG (Groupement de Recherches et d'Etudes des Génomes) and from the MESR and by the CNRS.

REFERENCES

1 Karp,P., Riley,M., Paley,S. and Pelligrini-Toole,A. (1996) *Nucleic Acids Res.*, **24**, 32–40.
2 On-line Mendelian Inheritance in Man (OMIM), a catalog of human genes and genetic disorders. McKusick,V.A. *et al.*, Johns Hopkins University. URL—http://www3.ncbi.nlm.nih.gov/omim/.
3 Fasano,L., Röder,L., Coré,N., Alexandre,E., Vola,C., Jacq,B. and Kerridge,S. (1991) *Cell*, **64**, 63–79.
4 Röder,L., Vola,C. and Kerridge,S. (1992) *Development*, **115**, 1017–1033.
5 Alexandre,E., Graba,Y., Fasano,L., Gallet,A., Perrin,L., De Zulueta,P., Pradel,J., Kerridge,S. and Jacq,B. (1996) *Mech. Dev.*, **59**, 191–204.
6 Bertuccioli,C., Fasano,L., Jun,S., Wang,S., Sheng,G. and Desplan,C. (1996) *Development*, **122**, 2673–2685.
7 Graba,Y., Aragnol,D., Laurenti,P., Garzino,V., Charmot,D., Berenger,H. and Pradel,J. (1992) *EMBO J.*, **11**, 3375–3384.
8 Gaul,U. and Jäckle,H. (1990) *Adv. Genet.*, **27**, 489–504.
9 Nüsslein-Volhard,C. and Wieschaus,E. (1980) *Nature*, **287**, 795–801.
10 Lewis,E.B. (1978) *Nature*, **276**, 565–570.
11 Abola,E.E., Bernstein,F.C. and Koetzle,T.F. (1988) In: *Computational molecular biology. Sources and methods for sequence analysis* (Lesk A.M., ed.), pp. 69–81, Oxford University Press, Oxford.
12 Benson,D.A., Boguski,M., Lipman,D.J. and Ostell,J. (1996) *Nucleic Acids Res.*, **24**, 1–5.
13 Rice,C.M., Fuchs,R., Higgins,D.G., Stoehr,P.J. and Cameron,G.N. (1993) *Nucleic Acids Res.*, **21**, 2967–2971.
14 George,D.G., Barker,W.C., Mewes,H-W., Pfeiffer,F. and Tsugita,A. (1996) *Nucleic Acids Res.*, **24**, 17–20.
15 Bairoch,A. and Apweiler,R. (1996) *Nucleic Acids Res.*, **24**, 21–25.
16 Wingender,E., Dietze,P., Karas,H. and Knüppel,R. (1996) *Nucleic Acids Res.*, **24**, 238–241.
17 The FlyBase Consortium (1996) *Nucleic Acids Res.*, **24**, 53–56.
18 Flyview, Janning,W. *et al.* A Drosophila Image Database. University of Münster. URL—http://pbio07.uni-muenster.de/.