Nucleic Acids Research, 1999, Vol. 27, No. 1 89–94

Grasping at molecular interactions and genetic networks in *Drosophila melanogaster* using FlyNets, an Internet database

Catherine Sanchez, Corinne Lachaize[1], Florence Janody, Bernard Bellon[2], Laurence Röder, Jérôme Euzenat[1], François Rechenmann[1] and Bernard Jacq*

Laboratoire de Génétique et Physiologie du Développement, IBDM, Parc Scientifique de Luminy, CNRS Case 907, 13288 Marseille Cedex 09, France, [1]INRIA Rhône-Alpes, 655, Avenue de l'Europe, 38330 Montbonnot St Martin, France and [2]Atelier de Bio-Informatique, Case 13, Université de Provence, 3 Place Victor Hugo, 13331 Marseille Cedex 03, France

Received October 2, 1998; Revised October 7, 1998; Accepted October 27, 1998

ABSTRACT

FlyNets (http://gifts.univ-mrs.fr/FlyNets/FlyNets_home_page.html) is a WWW database describing molecular interactions (protein–DNA, protein–RNA and protein–protein) in the fly *Drosophila melanogaster.* It is composed of two parts, as follows. (i) FlyNets-base is a specialized database which focuses on molecular interactions involved in *Drosophila* development. The information content of FlyNets-base is distributed among several specific lines arranged according to a GenBank-like format and grouped into five thematic zones to improve human readability. The FlyNets database achieves a high level of integration with other databases such as FlyBase, EMBL, GenBank and SWISS-PROT through numerous hyperlinks. (ii) FlyNets-list is a very simple and more general databank, the long-term goal of which is to report on any published molecular interaction occuring in the fly, giving direct web access to corresponding abstracts in Medline and in FlyBase. In the context of genome projects, databases describing molecular interactions and genetic networks will provide a link at the functional level between the genome, the proteome and the transcriptome worlds of different organisms. Interaction databases therefore aim at describing the contents, structure, function and behaviour of what we herein define as the interactome world.

INTRODUCTION

All known biological processes are controlled by direct and specific molecular interactions, involving DNA, RNA and proteins. These interactions form complex genetic interaction networks which are capable of responding to both external stimuli and stresses and also to internal changes occurring within components of the network. Being able to formally describe interactions and networks, to query and manipulate them, is slowly being recognized as an essential need for biologists studying gene regulation and function. From a more general point of view, integration of the structure and function of individual genes, RNAs and proteins with the knowledge of macromolecular interactions and networks in model organisms and in humans is an important step towards the construction of a unified and physiological view of the organism. Three major types of interactions, i.e., protein–DNA, protein–RNA and protein–protein interactions, account for the great majority of known biological macromolecular interactions. Several general databases exist for each of the three types of informational macromolecules (DNA, RNA, proteins), such as GenBank (1), EMBL (2) and DDBJ (3) databases for DNA and RNA sequences, SWISS-PROT (4) and PIR (5) databases for protein sequences and the PDB database (6) for molecular 3D structures. Many more specialized databases exist for specific families of genes, RNAs and proteins (see this issue of *Nucleic Acids Research* for an up-to-date collection of such databases). As has been more extensively discussed previously (9), data describing known specific molecular interactions between genes, RNA and proteins are under-represented in these databases and difficult to query.

Therefore, in order to adequately describe and study interactions, the development of new computer tools is necessary, among which that of powerful and up-to-date interaction databases is probably one of the most urgently needed. In 1995, we published on the web the GIF-DB database, a prototype for a *Drosophila* interactions database and, since then, other tools and databases have been proposed by our groups (7–9). More recently, other people have also developed databases which aim at describing interactions and networks involved in signal transduction pathways (10,11) or in transcriptional regulation (12–14). The yeast protein database YPD now also reports many protein–protein interactions (15).

In this paper, we describe the concepts, organization, contents and recent developments of FlyNets, a database which focuses on *Drosophila melanogaster* molecular interactions. To the best of

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

our knowledge, FlyNets is the only example where protein–DNA, protein–RNA and protein–protein interactions are described in the same database, using a unified description scheme.

BIOLOGICAL ASPECTS OF MOLECULAR INTERACTIONS AND GENETIC NETWORKS

Molecular and genetic interactions are classical biological concepts, which are sometimes mistaken. Molecular interactions imply a direct, physical contact between two molecules (for the purpose of this paper, only DNA, RNA and protein molecules will be considered). Many different experimental approaches have been developed in the last 20 years which can help to demonstrate that two molecules are interacting and that this interaction is biologically relevant.

The concept of genetic interaction is different: it is based on a genetic approach and is used to say that two genes are likely to be in the same biological pathway. Genetic interactions can be either direct interactions (as described above) or indirect ones, which are usually interpreted as a combination of several direct interactions, in which one or more intermediary gene(s) are propagating information from one studied gene to the other.

Understanding genetic interactions and describing them in terms of networks of direct biological interactions has long been, and still presently is, a great biological challenge. The concept of interaction network, although it has very often been used in the scientific literature, is still awaiting a precise and widely accepted definition. We view a genetic interaction network as a combination of molecular interactions, the understanding of the structure and function of which will be essential to face another biological challenge: the understanding of the relationships between genotypes and phenotypes.

It will also be essential in the future to consider interactions and regulatory networks at the level of a complete genome. One possible new way to look at genomes is to consider that a crucial aspect of their function is to code for products which are programmed to establish specific interactions. According to this view, the total number of genes of an organism is less important than the complete repertoire of interactions potentially encoded by its genome (the interactome). Indeed, a small difference in the number of genes between two organisms could be sufficient to cause a large increase in the number of interactions (and hence a larger organizational complexity of the organism), provided that the 'new' genes code for proteins with a large interaction potential. It is our belief that important progress in our present knowledge of interaction networks and of the interactome are essential to understand how gene functions and regulations are integrated at the level of an organism.

One essential aspect of this study of gene regulatory networks lies at the basic level of collecting and describing molecular interactions and genetic networks. In order to do so, precise definitions and classifications of molecular interactions are needed. Our working definition for a gene molecular interaction is the following: 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). Among the six different molecular interaction types which could then theoretically be considered we have focused on three major types of interactions only, which are by far the most documented ones, whatever the organism being considered: protein–DNA interactions, protein–RNA interactions and protein–protein interactions.

THE FlyNets DATABASE

Purpose and leading concepts of FlyNets

We are interested in the process of pattern formation in *Drosophila* and in understanding the basis of specific identity acquisition by the different body parts (16–18). In order to help us in the description of specific developmental interactions, the GIF-DB (Gene Interactions in the Fly Database) and then the FlyNets databases have been developed (7,9). FlyNets-base is now officially the successor of GIF-DB which will no longer be accessible after March 1, 1999. FlyNets has the same general objectives as GIF-DB: it is a WWW database which aims at providing a repository for data on gene interactions involved in *Drosophila* development and the regulatory networks in which they are involved. FlyNets also shares with GIF-DB the main leading concepts and specific goals which are: (i) to propose a simple but efficient way to represent the various and complex knowledge on molecular interactions in *Drosophila* development; (ii) to attain a high level of integration with other databases; (iii) to classify all molecular interactions in one of the three major interaction types (protein–DNA, protein–RNA or protein–protein interactions); and (iv) to define a generic mode of interaction representation which could potentially be used for the description of nearly any gene interaction, whatever the biological process and the organism in which they occur may be.

In line with previous biological considerations, all molecular interactions in the FlyNets database are described as binary interactions (i.e., interactions occurring between two molecular partners). Any complex interaction which involves more than two partners (interaction between a DNA sequence and several different transcription factors, or between several proteins into a multimeric complex, for instance) will therefore be described using several binary interactions. Since analysis of genetic and molecular interactions usually involves the study of two entities at a time, this binary point of view also fits well with our current experimental approaches.

The differences between GIF-DB and FlyNets-base are at the level of the entry format and will be discussed in the next sub-section.

Database organisation and entry format

The GIF-DB and FlyNets databases are collections of hypertext file entries in which each entry belongs to one of the interaction classes defined above and describes an interaction between two molecular partners. FlyNets-base is an annotated database which does not contain any primary raw data. Scientific data concerning specific interactions are extracted from the literature, verified, compiled and entered into a relational database built using the 4D DBMS (ACI inc.) on a MacIntosh computer (F.Horn, M.Imbert and B.Jacq, unpublished). The HTML files constituting FlyNets are then automatically generated from this database.

Information in FlyNets entries has been arranged into a 'GenBank-like' model format. Each entry is composed of lines and 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. For the sake of clarity, the different linetypes in a FlyNets entry have been arranged into five zones: the ENTRY

zone, the EFFECTOR zone, the TARGET zone, the INTERACTION zone and the REFERENCES zone. Many of the FlyNets linetypes are the same as those in GIF-DB, and this is also true for the conventions adopted in the line contents (and described in the on-line FlyNets-primer document), for the line length (80 characters) and for the format of the references. Some of the lines accept only a controlled vocabulary (e.g., the protein localization, protein function or interaction type lines), whereas others, like the different comment lines, accept free text. Wherever possible, symbols and nomenclature which are supposed to be familiar to drosophilists, geneticists, biochemists and molecular biologists are used to describe the interactions and some conventions used in FlyBase, the genetic and molecular *Drosophila* database (19) have been followed. Finally, many of the lines in FlyNets directly point towards external databases such as FlyBase, EMBL, GenBank or SWISS-PROT through numerous hyperlinks.

The two main differences between the former GIF-DB database and its successor FlyNets are the number of linetypes supported and the format of the line headers. A total of 31 linetypes (instead of 40 in GIF-DB) are presently supported. The comment line in FlyNets now regroups data which were present in several different comment lines in GIF-DB. This line is organised in a similar way to the CC line in SWISS-PROT, in which different types of comments are arranged in as many sub-comments. The second difference with GIF-DB is the format of linetype headers: each line in a GIF-DB entry started with a two-character line code indicating the type of information contained in the line (as is the case for EMBL and SWISS-PROT). Since several users of GIF-DB found it difficult to memorise the signification of a two-letter code for 40 linetypes, we have decided to adopt a different convention in FlyNets. The line headers now have a GenBank-like format and are explicit words or group of words with a maximum length of 20 characters (e.g., identificator, creation date, target function, authors, ...). More details on the database general organization, entry format and the different linetypes can be found on the on-line version of FlyNets. Version 2.0 of FlyNets-base (January 1999) contains 80 detailed interactions. Until November 1st, 1998, FlyNets-base was accessible from the list of entries only. Another way of accessing FlyNets data, through the use of a powerful search program is now available (see recent developments below).

On February 1st, 1998, FlyNets-interactilist, a companion to FlyNets (and now called FlyNets-list) was been loaded on the GIFTS Server. FlyNets-list is aimed at providing a list of known molecular interactions in *Drosophila*, without any added information, except hyperlinks towards corresponding bibliographic references in Medline and Flybase. Version 1.0 of FlyNets-Interactilist (February 1998) contained 130 interactions and 170 associated bibliographic reference links; version 2.0 (July 1998) contained 175 interactions and 280 bibliographic reference links. The last update (FlyNets-list version 3.0, December 1998) contains 210 interactions and 350 references. Version 4.0 is scheduled for February 1999 and will contain ~240 interactions and 400 reference links. In the long term, all developmental interactions in FlyNets-list will become FlyNets-base entries. An example of interactions in FlyNets-list is given in Figure 1.

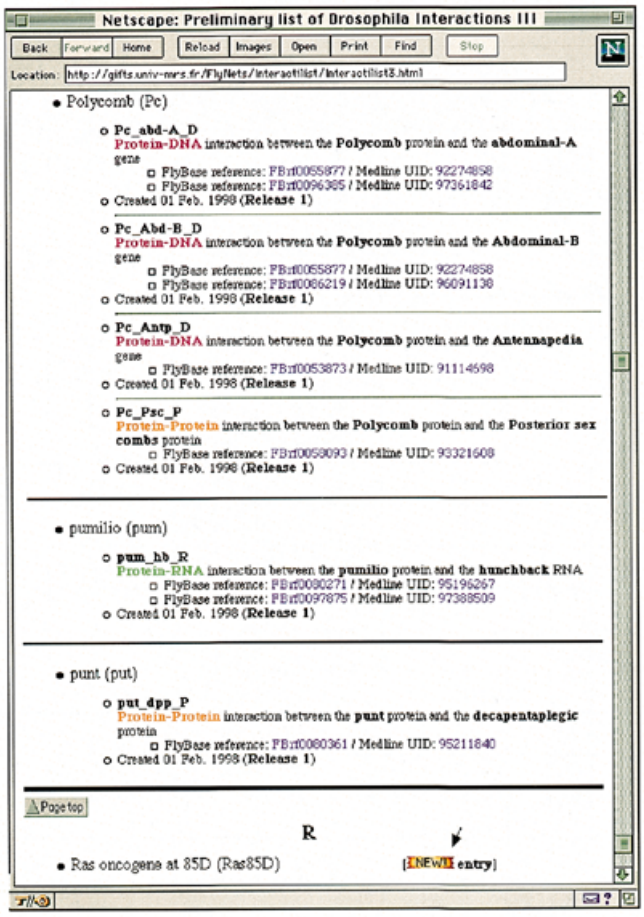

Figure 1. A portion of a page from Flynets-list. All interactions are alphabetically ranked according to the name of the effector. For each interaction the abbreviation, the descriptor and type of the interaction, the bibliographic links in Flybase and Medline and the creation date are given. A color code allows the user to grasp at a glance the molecular type of the interaction (pink: protein–DNA, green: protein–RNA and yellow: protein–protein). New entries in the latest release are flagged by a logo (arrow).

Recent developments

A powerful query search program has now been introduced in FlyNets. It allows searching for the occurrence of an ASCII character string entered by the user. The search can be performed either on the text of the entire database or in any one of the 31 different data lines of all entries. A complete or partial word-matching option is available (Fig. 2).

Many interactions in the database are linked together in the sense that the target gene for a given protein–DNA interaction (for instance) codes for a protein which is itself the effector of another protein–DNA, protein–RNA or protein–protein interaction. Therefore, many genetic networks can be constructed using interactions described in FlyNets-base and/or FlyNets-list. In order to obtain a graphical display of these interactions, we are developing an interactive graphical network editor which will soon be interfaced with FlyNets. At the moment, interactions are manually selected from the entire list of interactions and graphically displayed, using a color code for distinguishing protein–protein, protein–DNA and protein–RNA interactions. Since genes (proteins) and interactions are defined as graphical objects, the user can then manually arrange interactions in order to highlight one peculiar feature. A typical display from this editor is given in Figure 3 where genes, originally displayed along

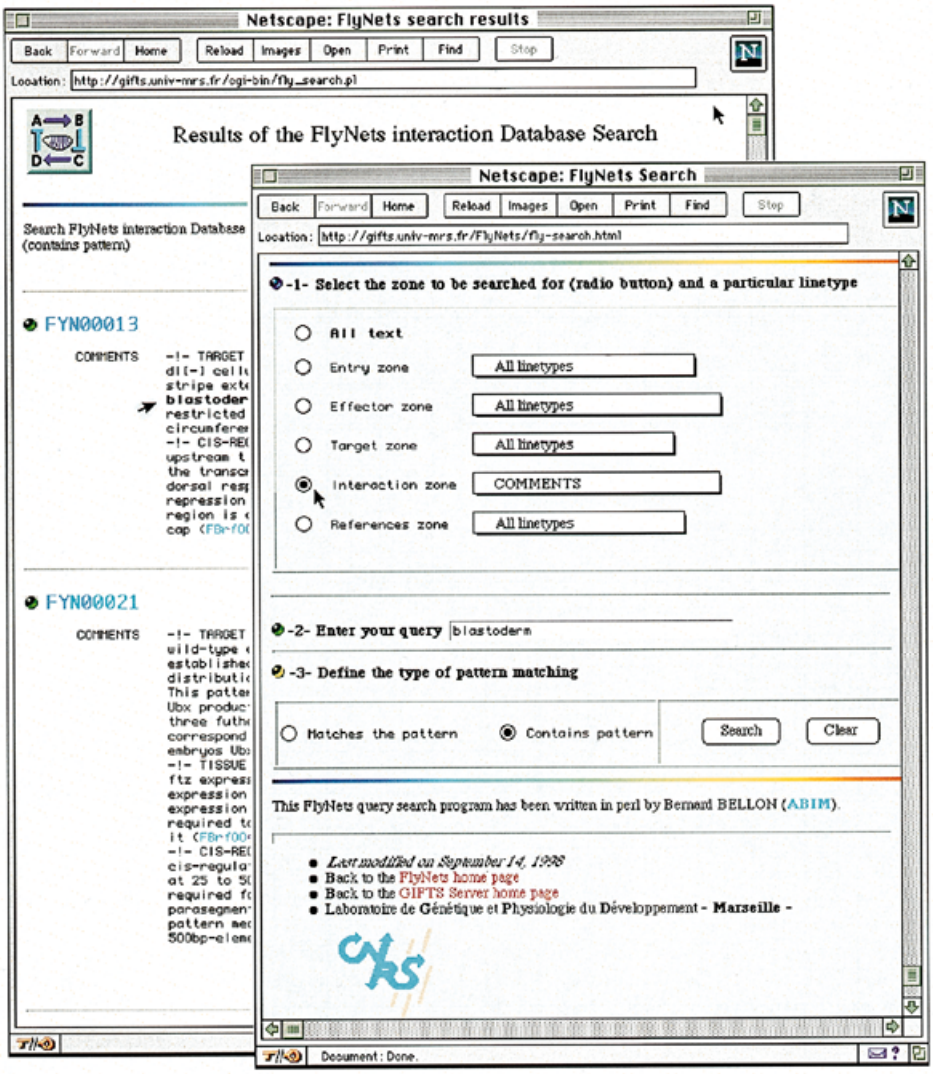

Figure 2. An example of a FlyNets-base query search and result. In the query window (front panel), the user has selected the comment linetype in the interaction zone and searches for sentences containing the word 'blastoderm'. The result is presented in the back panel window. All Flynets entries matching the query are listed and only comment lines are displayed. The full-length Flynets entries can be directly accessed through a hyperlink (in blue) and the queried term is displayed in bold (arrow).

a circular circumference, have now been manually grouped according to their biological function.

CONCLUSIONS AND FUTURE PROSPECTS

The different genome projects are presently providing us with extensive catalogs of genes and proteins for several model organisms. After the genomes from the most extensively studied organisms (both prokaryotic and eukaryotic) have been deciphered, it will be essential to describe how and with which molecular partners the different genes, RNA and proteins from an organism establish specific interactions, a knowledge which cannot be simply deduced from their sequences or structures. In what is now called the 'post-genome' era, both new experimental methods and new bioinformatics concepts and tools are needed to gradually paint the complicated picture of biological pathways. On one hand, new large-scale experimental methods such as exhaustive two-hybrid screens (20) and DNA micro-arrays (21) are likely to bring new high-throughput data on protein–protein and genetic/physiological interactions, respectively. On the other hand, going back to the scientific literature to extract specific knowledge on the thousands of already known interactions will also be of crucial importance in order to integrate new knowledge with known results not yet explicitly described in databases. In this respect, we are presently implementing new algorithms to help extract pertinent information on interactions from texts, relying either on advanced linguistic tools, completed with object-oriented knowledge modeling capabilities (Proux *et al.*, submitted) or on statistical methods applied to group of words (Pillet *et al.*, in preparation).

Once lists of interactions have been created, relying on experimental data and/or text analysis methods as described above, interaction databases are necessary to describe and query this functional knowledge. Databases such as FlyNets provide a simple and straightforward way to make functional links between specific entries from different molecular databases. Such func-

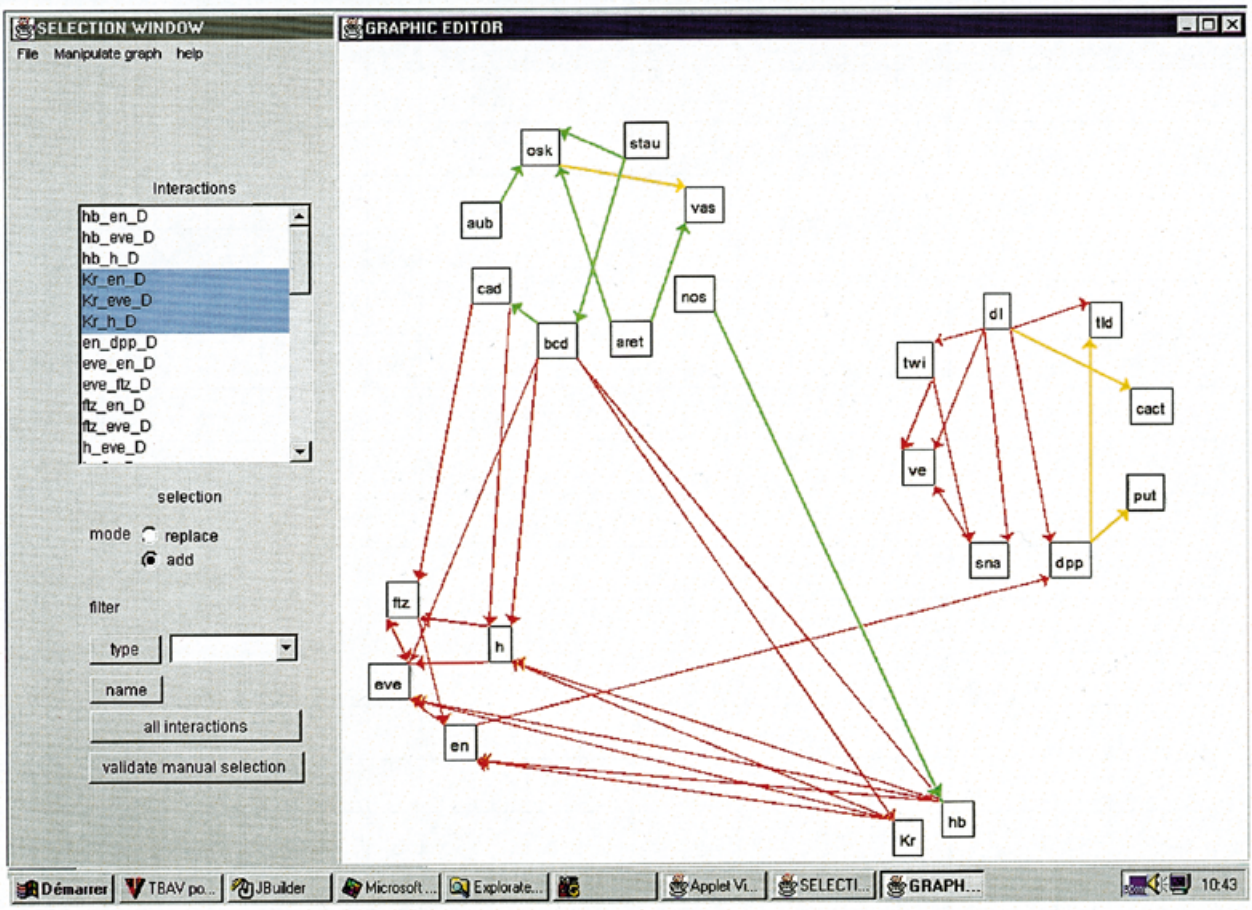

Figure 3. Graphic display of Flynets-base interactions. In the selection window (left panel) the user chooses from the complete list of interactions those which will be displayed. The resulting interaction network is then drawn in the graph editor window (right panel). The abbreviated name of the gene is presented in boxes. Two interacting partners are linked by an arrow, the color of which corresponds to the defined interaction classes (see Fig. 1 legend). In this example, genes have been manually grouped by the user according to their biological function. Maternal, dorso-ventral, pair-rule and gap genes are respectively on the top left, top right, bottom left and bottom right parts of the graph.

tional links are a useful complement to the structural links (present as database cross-references) already existing between many EMBL (or Genbank) entries and their SWISS-PROT (or PIR-International) corresponding translational products.

Within the next few years, we plan to offer new possibilities in FlyNets. This will be done first through the addition of a few new linetypes and the adjunction of hyperlinks towards other databases such as the Inter*active* Fly, a cyberspace guide to *Drosophila* genes and their roles in development (T.Brody; http://sdb.bio.purdue.edu/fly/aimain/1aahome.htm) or TRANSFAC (12). Second, many new functionalities will be added to the interactions and networks graphical editor. New query tools will be developed, allowing the user to select interactions according to the biochemical structure, the functional class or the biological role of the genes/proteins. It will also be possible to declare that an ordered set of interactions define one regulatory pathway which will be stored under a specific name (the wingless pathway, the hedgehog pathway, the Toll pathway, ... just to cite a few). When looking at a large number of interactions, such biological pathways will then be graphically highlighted, in order to analyze the relationships between different pathways.

Part of our data on interactions has now been included in KNIFE, an object-oriented knowledge base on interactions presently under development (8). Such a base will allow us to better cope with complex notions such as developmental stages, tissue-restricted gene expression or dynamic multi-protein complexes which cannot be easily described in the context of a collection of flat hypertext files such as FlyNets. Our aim is to integrate all FlyNets data into KNIFE which will then become the computer tool through which FlyNets data will be entered and queried. Furthermore, the future integration in KNIFE of data on mouse developmental genes orthologous to *Drosophila* genes described in FlyNets will allow us to study interaction networks in an evolutionary perspective and analyze to what extent homologous genes are working through homologous regulatory pathways.

INTERACTIVE WWW ACCESS AND QUOTING FlyNets

FlyNets can be accessed using the World Wide Web through the GIFTS (Gene Interaction in the Fly Transworld Server) WWW server in Marseille. The URL of the GIFTS Server is

http://gifts.univ-mrs.fr/GIFTS_home_page.html . Using this server, one can also access SOS-DGDB (9), a collection of annotated *Drosophila* gene sequences, in which binding sites for regulatory proteins are directly visible on the DNA primary sequence and hyperlinked to TRANSFAC database entries.

In addition to giving access to these databases, the GIFTS server also provides services such as GIN, a series of annotated pages to help navigate on the Internet and BLASTula, a specialized service giving an integrated access to more than 60 different BLAST analysis sites in the world. Many of these BLAST servers operate on collections of new DNA sequences from the different genome projects which are not yet integrated in the EMBL, GenBank and DDBJ general databases. Using BLASTula therefore augments the probability of finding significant hits with a queried sequence.

If you use FlyNets-base and/or FlyNets-list as tools in your published research work, please cite this paper. Comments and enquiries about FlyNets are welcome and should be sent to Bernard Jacq (Email: jacq@lgpd.univ-mrs.fr)

ACKNOWLEDGEMENTS

We would like to thank our colleagues in the LGPD and INRIA Rhône-Alpes for helpul discusssions on computer and biological issues about interactions. We are grateful to Marie Imbert and Florence Horn for their help in the modification of the relational database program managing FlyNets interactions. This work has been supported by a CNRS ‘programme genome’ grant to B.J.

REFERENCES

1 Benson,D.A., Boguski,M.S., Lipman,D.J, Ostell,J. and Ouellette,B.F.F., (1998) *Nucleic Acids Res.*, **26**, 1–7.
2 Stoesser,G., Moseley,M.A., Sleep,J., McGowran,M., Garcia-Pastor,M. and Sterk,P. (1998) *Nucleic Acids Res.*, **26**, 8–15.
3 Tateno,Y., Fukami-Kobayashi,K., Miyazaki,S., Sugawara,H. and Gojobori,T. (1998) *Nucleic Acids Res.*, **26**, 16–20.
4 Bairoch,A. and Apweiler,R. (1998) *Nucleic Acids Res.*, **26**, 38–42.
5 Barker,W.C., Garavelli,J.S., Haft,D.H., Hunt,L.T., Marzec,C.R., Orcutt,B.C., Srinivasarao,G.Y., Yeh,L.-S.L., Ledley,R.S., Mewes,H.-W., Pfeiffer,F. and Tsugita,A. (1998) *Nucleic Acids Res.*, **26**, 27–32.
6 Abola,E.E., Bernstein,F.C. and Koetzle,T.F. (1988) In Lesk,A.M. (ed.), *Computational Molecular Biology. Sources and Methods for Sequence Analysis*. Oxford University Press, Oxford, UK, pp. 69–81.
7 Jacq,B., Horn,F., Janody,F., Gompel,N., Serralbo,O., Mohr,E., Leroy,C., Bellon,B., Fasano,L., Laurenti,P. and Röder,L. (1997) *Nucleic Acids Res.*, **25**, 67–71.
8 Euzenat,J., Chemla,C. and Jacq,B. (1997) In *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, CA, pp. 108–119.
9 Mohr,E., Horn,F., Janody,F., Sanchez,C., Pillet,V., Bellon,B., Röder,L. and Jacq,B. (1998) *Nucleic Acids Res.*, **26**, 89–93.
10 Goto,S., Bono,H., Ogata,H., Fujibuchi,W., Nishioka,T. and Kanehisa,M. (1996) *Proceedings of the Pacific Symposium on Biocomputing*, pp. 175–186.
11 Igarashi,T. and Kaminuma,T. (1997) *Proceedings of the Pacific Symposium on Biocomputing*, pp. 187–197.
12 Heinemeyer,T., Wingender,E., Reuter,I., Hermjakob,H., Kel,A.E., Kel,O.V., Ignatieva,E.V., Ananko,E.A., Podkolodnaya,O.A., Kolpakov,F.A., Podkolodny,N.L. and Kolchanov,N.A. (1998) *Nucleic Acids Res.*, **26**, 362–367.
13 Thieffry,D., Salgado,H., Huerta,A.M. and Collado-Vides,J. (1998) *Bioinformatics*, **14**, 391–400.
14 Kolpakov,F.A., Ananko,E.A., Kolesov,G.B. and Kolchanov,N.A. (1998) *Bioinformatics*, **14**, 529–537.
15 Hodges,P.E., Payne,W.E. and Garrels,J.I. (1998) *Nucleic Acids Res.*, **26**, 68–72.
16 Fasano,L., Röder,L., Coré,N., Alexandre,E., Vola,C., Jacq,B. and Kerridge,S. (1991) *Cell*, **64**, 63–79.
17 Röder,L., Vola,C. and Kerridge,S. (1992) *Development*, **115**, 1017–1033.
18 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.
19 FlyBase Consortium (1998) *Nucleic Acids Res.*, **26**, 85–88.
20 Fromont-Racine,M., Rain,J.C. and Legrain,P. (1997) *Nature Genet.*, **16**, 277–282.
21 Lashkari,D.A., DeRisi,J.L., McCusker,J.H., Namath,A.F., Gentile,C., Hwang,S.Y., Brown,P.O. and Davis,R.W. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 13057–13062.