GPCRDB: an information system for G protein-coupled receptors

F. Horn, J. Weare, M. W. Beukers[1], S. Hörsch[2], A. Bairoch[3], W. Chen[4], Ø. Edvardsen[5], F. Campagne[6] and G. Vriend*

BIOcomputing, EMBL, Heidelberg, Germany, [1]LACDR, Farmacochemie, LEIDEN, The Netherlands, [2]EBI, Hinxton, Cambridge, UK, [3]Department of Medical Biochemistry, University of Geneva, Geneva, Switzerland, [4]GMD, Darmstadt, Germany, [5]Institute of Medical Biology, University of Tromsø, Norway and [6]Laboratoire de Chimie Théorique, Nancy, France

Received September 22, 1997; Revised and Accepted October 24, 1997

ABSTRACT

The GPCRDB is a G protein-coupled receptor (GPCR) database system aimed at the collection and dissemination of GPCR related data. It holds sequences, mutant data and ligand binding constants as primary (experimental) data. Computationally derived data such as multiple sequence alignments, three dimensional models, phylogenetic trees and two dimensional visualization tools are added to enhance the database's usefulness. The GPCRDB is an EU sponsored project aimed at building a generic molecular class specific database capable of dealing with highly heterogeneous data. GPCRs were chosen as test molecules because of their enormous importance for medical sciences and due to the availability of so much highly heterogeneous data. The GPCRDB is available via the WWW at http://www.gpcr.org/7tm

INTRODUCTION

G protein-coupled receptors (GPCRs) consist of a single protein chain that crosses the membrane seven times. These, presumably α-helical, transmembrane regions are probably arranged with similarity (1) to bacteriorhodopsin. Except for low resolution electron diffraction studies of frog and bovine rhodopsins (2,3) and NMR structures of fragments of loops of the bovine rhodopsin receptor (4–6) not much solid structural information is available. GPCRs are of enormous importance for the pharmaceutical industry because 52% of all existing medicines act on a GPCR (7). This explains why so many three dimensional models of GPCRs have been built. Most models are based on the atomic coordinates of the bacteriorhodopsin structure. In all of these modeling studies (bio)chemical and pharmacological data had to be used to refine the models. Clearly, mutation data and structure–affinity relationships (SARs) are essential for modeling studies aimed at the analysis of receptor–ligand interactions.

It is also possible to arrive at detailed structural knowledge using sequence analysis techniques combined with extensive data mining rather than via the path of building models. Correlated mutation studies, for example, have shown great potential for the determination of inter molecular contacts between GPCRs and their G proteins or ligands (9,10), for drug design purposes, for analyzing the role of olfactory receptors in the organization of the olfactorial nerve system (8) etc. None of these correlated mutation studies could have been performed without the availability of large amounts of experimental data.

Due to the lack of high resolution structural data, theoretical research on GPCRs relies on the availability and easy accessibility of all available data in an information system that allows for the four basic data dissemination functions: browsing, retrieval, querying and inferencing. Additionally, this information system should provide tools for data gathering, data input, data validation and data annotation. We will first discuss the data and then the four dissemination facilities (see also Fig. 1). In all cases we will discuss the present situation and the plans for the near future.

DATA

Data types introduction

The GPCRDB holds three kinds of experimental data: sequences, mutation data and ligand binding data. All three data types have their own specific problems. Sequences can be wrong, truncated, or can occur in several alternative translations. The interpretation of mutation studies and ligand bindings studies depend strongly on experimental conditions such as cell type, stable/transient transfection, density of e.g. receptors, second messenger assay, and the kinds of ligands used.

Additional data (types) can easily be incorporated in the GPCRDB. We will, however, only enter new data types if a certain degree of completeness can be reached, and if a mechanism for updating exists. We would like to add data about, for example, the preferred G protein, receptor localization, second messenger, etc., but at present it does not seem likely that a high degree of completeness can be reached for these kinds of data.

The number of experimental data types is limited, but there are no limits to the number of data types that can be derived computationally. It is therefore important to think about the questions that the GPCRDB should help answer when adding

*To whom correspondence should be addressed at: Meyerhofstrasse 1, 69117 Heidelberg, Germany. Tel: +49 6221 387473; Fax: +49 6221 387517; Email: vriend@embl-heidelberg.de

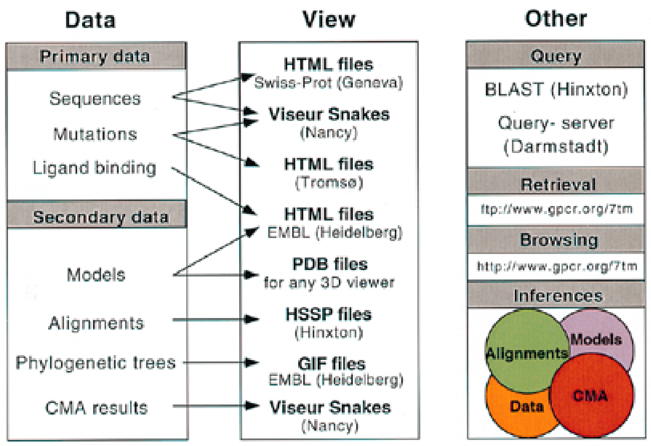


Figure 1. Schematic representation of the GPCRDB organisation. Data is shown at the left, the viewing methods in the middle and the database facilities at the right. Cities in brackets indicate the original location of the data. Arrows indicate the viewing methods available for each data type. The coloured circles represent the interplay between the data and the methods that can lead to discoveries.

more and more computational data. At present the central computational data type is multiple sequence alignments (for which a method was designed that is dedicated for GPCRs; 11). Phylogenetic trees and correlated mutation analyses (CMA; 11) are derived from these multiple sequence alignments.

Visualization of data is important for the human users of the system. In the GPCR field two dimensional representations by so-called snake like diagrams (or snakes for short) are commonly used to visually combine a sequence with other types of information such as three dimensional localization, mutation results, ligand binding or biochemical studies. In the GPCRDB there is a snake for every sequence, and these diagrams are hyperlinked to the most important related data types.

It is not a very big job to keep the computationally derived data complete, but the extraction of the underlying experimental data from literature is such an overwhelming task that in the long run it cannot be done without the help of the experimental scientists around the world. Convincing the community that there is mutual benefit for individuals to infrequently invest small amounts of time entering data might be a more daunting task than all other aspects of the GPCRDB project. Moreover, the direct input of data by the experimentalists allows the incorporation of non-published data in the GPCRDB. The inclusion of such data will prevent researchers from reinventing the wheel.

Primary data

Sequences. At present, all sequence data is extracted from the SWISS-PROT database (12). This has the advantage that we get well annotated sequences from expert database curators. The disadvantage is that the GPCRDB sequence data at present is not yet as complete and as up to date as possible. SWISS-PROT maintains lists of sequences for certain classes for which niche databases such as the GPCRDB exist. In the future the SWISS-PROT data will be augmented with the results of sequence searches in other sequence databases. The SWISS-PROT sequence files will, however, stay at the heart of the GPCRDB sequence efforts because of their high quality.

In the September 1997 release of the GPCRDB, more than 800 GPCR sequences are available, divided into five major and 151 minor sub-families.

Sequence fragments are deleterious for most computational purposes but they do hold useful information. At present (September 1997) we have stored 74 sequence fragments, extracted from SWISS-PROT. In the near future, putative GPCR ESTs (Expressed Sequence Tags) will be added.

Mutation data. The GRAP database (13) and its derivative, TinyGRAP, holds information for about 3900 point mutations. Some data on chimeric, deletion and insertion mutations is also available in the GRAP database and the planning is to include more of such mutations in the near future. This information is accessible directly (14), from the Viseur site (15) and from the GPCRDB. In the GPCRDB this data is available in several forms, the most prominent of which is hyperlinking from the snakes. Presently, GRAP focuses on class A receptors but an input form to add mutant data for all classes is available, and by the time this article goes to press, the first 100 class B mutations will have been entered. We strongly encourage the experimental scientists in the GPCR field to enter their mutation data into the GRAP database via the WWW input form available from the GPCRDB pages.

Ligand binding data. Ligand binding data was obtained from P.Seeman (16). This very impressive collection of drug dissociation constants was manually extracted from the literature. Seeman collected data for neuroreceptors and transporters. About 12 000 dissociation constants are available for 10 GPCR families. The heterogeneity of the experimental conditions make it hard to use this data for query purposes because it is not possible to determine the value of the numbers without human interpretation. However, the data is very useful for browsing purposes. The data was originally stored in tables that were sorted by receptor type, but in the near future the data will be stored internally in a data structure that will allow for alternative visualization methods such as sorting by ligand, and for hyperlinking to facilitate more natural browsing behavior by the users.

Secondary data

Multiple sequence alignments. Multiple sequence alignments are performed with WHAT IF (17) as described by Oliveira *et al.* (11). These alignments are made for whole families, but also for sub-families, sub-sub-families, etc. The alignments are presented as HSSP files (18) and as MSF files (19) so that the user can choose between two standard output formats one of which displays the sequences horizontally (MSF), and one vertically (HSSP). Although seemingly trivial, interactions with users made clear that this freedom of choice is rather important.

Throughout the GPCRDB we have used the residue numbering scheme suggested by Oliveira *et al.* (11). In this numbering scheme the residues are numbered such that 100s digits indicates the helix number, and the most conserved residue in every helix has a round number.

Alternative alignments will be incorporated if submitted to the GPCRDB.

Phylogenetic trees. Phylogenetic trees are a good visualization tool for relationships between sequences in a family. This

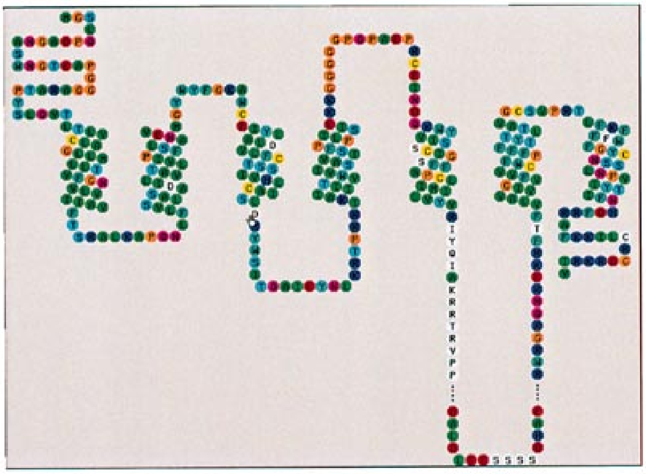


Figure 2. Snake like diagram of an adrenergic receptor. Residues are coloured based on their biochemical nature except white-coloured positions which are hyperlinked to mutant data.

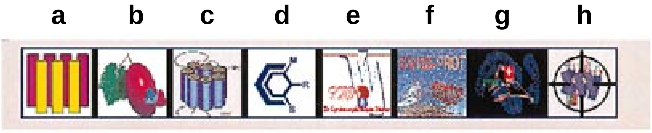


Figure 3. Logos representing the informal collaboration of GPCR databases. (**a**) The GPCRDB maintained at the EMBL. (**b**) The GRAP mutant database contains data on mutants of family A GPCRs maintained at the University of Tromsø by Ø.Edvardsen and K.Kristiansen. (**c**) The ORDB database holds sequences of olfactory receptors proteins. It contains public and private sections which provide tools for investigators to analyse the functions of this very large gene family of GPCRs. It is maintained at the Yale School of Medicine. (**d**) The Molecular Recognition Section. One of their projects is a database of GPCR related information including alignments and mutation analysis by A.M.van Rhee, NIH, Bethesda. (**e**) The GCRDb is maintained by F.Kolakowski at the University of Texas, Health Science Center at San Antonio. (**f**) The SWISS-PROT database: the annotated protein sequence database maintained by A.Bairoch at the University of Geneva, Geneva, Switzerland. (**g**) The Swiss-Model 7TM Interface for the modelling of the helices of 7TM receptors maintained by M.C.Peitsch at the Geneva Biomedical Research Institute, Geneva, Switzerland. (**h**) The Viseur program for the visualisation, management and integration of GPCR related information: maintained by F.Champagne at the Laboratoire de Chimie Théorique, Nancy, France.

information can help in answering several different kinds of questions, e.g., questions related to ligand design. Making phylogenetic trees is a whole science in itself. The WHAT IF program uses a neighbor joining algorithm. Although the resultant phylogenetic trees represent, as good as possible, the pairwise identities between the sequences rather than their evolutionary relationships, there is a striking resemblance with phylogenetic trees based on an accepted-mutation parsimony method (20). The phylogenetic trees presently available in the GPCRDB were made of manually selected representative subsets of sequences. We envisage producing these trees automatically, in parallel with the multiple sequence alignments.

Correlated mutation data. Correlated mutation analysis is a computational method to identify pairs of sequence positions that remained conserved or mutated in tandem during evolution. The idea behind the search for such pairs of residues is that when a mutation occurs at a functionally important site, the protein either becomes non-functional or may acquire its original or a different function due to a compensatory mutation at another position. Residues detected by the CMA method are often involved in intermolecular interactions (between ligands and receptors or G-proteins and receptors) (9,10). Although a detailed explanation for this phenomenon is beyond the scope of this article, it must be clear that the automatic detection of 'important' residues is a relevant aspect of the GPCRDB effort.

Snake like diagrams. Experimentalists in the GPCR field prefer to represent their data using snakes (Fig. 2). The Viseur (15) program can automatically generate snakes and hyperlink them to other types of information. Snakes are used in the GPCRDB to represent two kinds of data. One set of snakes is hyperlinked to the TinyGRAP mutant database. The second set is used to indicate the location of residues detected in the CMA analyses. This second set is hyperlinked to the corresponding HSSP alignment files.

Three dimensional models. The GPCRDB server holds atomic coordinates of 3D models of GPCRs. Different modelers used different alignments and different modeling techniques to build these models, and consequently a wide variety of models have been proposed. As it is at present not possible to decide which models are right and which are wrong, we have decided to store every suggested model in the GPCRDB. The models are grouped per depositor. For each model one can either download the coordinates or view them using a WWW helper application like Rasmol (21).

Other data

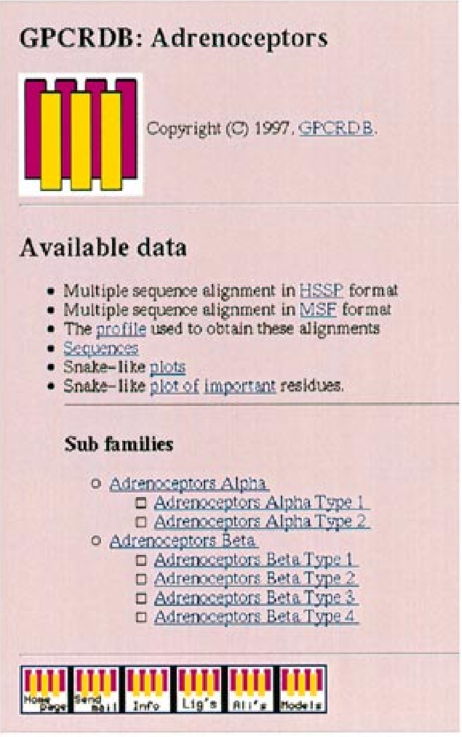
The GPCRDB additionally contains pointers to other GPCR related databases (Fig. 3). General information about GPCRs, articles about the GPCRDB or elements of its contents, lists of GPCR specialists addresses, pointers to external pages of different levels of relevance for GPCR research (information resources, articles, group pages, GPCR related diseases, etc.) are available. The SWISS-PROT files allow for navigation to other databases like Medline (22), OMIM (23), EMBL (24), PIR (25), Prosite (26) and several organism specific databases such as FlyBase (27) for *Drosophila* sequences. The mutation data in GRAP is hyperlinked to Medline and OMIM.

DISSEMINATION FACILITIES

GPCRDB has been conceived to provide fast and easy access to all information related to GPCRs. It should be an information tool that makes it easier for the user to think about GPCRs, and it should make suggestions for future research. For this purposes we have implemented, (and are still implementing) the four basic information system tools: browsing, retrieval, query and inferencing.

Browsing

The GPCRDB organization is based on the pharmacological classification of receptors and access to the data is obtained via a hierarchical list of known families in agreement with this classification. For one specific family, one can access the individual sequences, the multiple alignments, the profile used to perform the latter, the snakes and a phylogenetic tree. Each type of data is displayed in a WWW page with hyperlinks to other data



GPCRDB: Adrenoceptors

Copyright (C) 1997, GPCRDB.

Available data

- Multiple sequence alignment in HSSP format
- Multiple sequence alignment in MSF format
- The profile used to obtain these alignments
- Sequences
- Snake-like plots
- Snake-like plot of important residues.

Sub families

- ○ Adrenoceptors Alpha
  - □ Adrenoceptors Alpha Type 1
  - □ Adrenoceptors Alpha Type 2
- ○ Adrenoceptors Beta
  - □ Adrenoceptors Beta Type 1
  - □ Adrenoceptors Beta Type 2
  - □ Adrenoceptors Beta Type 3
  - □ Adrenoceptors Beta Type 4

Figure 4. Example of one of the 151 subfamilies specific pages. All underlined words and icons are hyperlinked for navigation through the GPCRDB.

where appropriate. Figure 4 shows the WWW page for the adrenergic receptor class as an example of the data organization.

Retrieval

Often a user wants to work on certain data at home, independent of the GPCRDB environment. Therefore most data can be retrieved in its native form using the ‘save as’ option of the WWW browsers or via anonymous FTP from www.gpcr.org, data being stored in the /7tm directory.

Query

A BLAST server allows the user to scan one GPCR sequence, or the fragment of it against all GPCRDB sequences. This can give for new sequences an impression about the family they belong too.

A query system is under development. In due time it will answer complicated questions such as, for example: ‘are there any mutant or CMA data for position 340 of the α2 adrenergic receptors?’ The current version supports simple sequence queries by specifying query conditions on specific fields using pattern matching by means of regular expressions that can be freely combined using the logical connectors ‘and, ‘or’ and ‘not’. It will be soon possible to combine these typical database queries with less exact ones such as the sequence similarity defined by FASTA. This query facility is accessed via a WWW interface that performs conversion of the users input to the query interface of the underlying database system and a formatted presentation of query results.

Inferences

Lacking high resolution structure data, correlated mutation analyses are the most powerful tool available to date for the computational discovery of novel facts about GPCRs. The CMA clearly provide a powerful inference engine. At the cost of little CPU time, the important residue positions are selected from among the tens of thousands of residues in each alignment. So far we only calculate these residues and make them available for browsing purposes by displaying them as snakes with the appropriate hyperlinks. We will try to combine this information with experimental data (i.e. mutation data and ligand binding data) to strengthen the conclusions that can be drawn and to make it easier for the GPCRDB users to investigate their relevance.

DISCUSSION

The GPCRDB is now officially one year old. In this first year we have created a computer infrastructure that flexibly allows for extension with new data types, experimental as well as computational. Many improvements can still be made. These improvements will be directed along four lines: (i) addition of new experimental data types (ligand information, preferred G protein, cellular localization, secondary messenger, disease pattern, chimera data); (ii) addition of new theoretical data types (docked ligands, codon usage); (iii) addition of hyperlinks to other databases (mouse knock-out database, genetic diseases, etc.); (iv) user-friendliness. It is envisaged that the GPCRDB will function in the same way as an encyclopedia. One opens it with the aim of answering a question, but multiple good ideas later one has forgotten what the original question was. In other words, the GPCRDB should be more than intuitive—suggestive.

The most complicated aspect of the GPCRDB effort is the actual incorporation of data. The vast majority of all useful data is stored in articles rather than in computer readable form. We are providing tools to enter data into the database, but convincing experimentalists that it is in their own interest to participate with their data for usage by the community is another important aspect.

USAGE AND AVAILABILITY

The GPCRDB is accessible from http://www.gpcr.org/7tm . The underlying data files (alignments, models, etc.) can be downloaded by anonymous FTP from www.gpcr.org/7tm. Access to the GPCRDB is free for academic and industrial scientists. Industries can download the entire GPCRDB for in-house usage from the same FTP site. In the first year of its existence the GPCRDB has been ‘visited’ on average 250 times per week.

ACKNOWLEDGEMENTS

The GPCRDB is an EU sponsored project (PL 950224). We thank K.Aberer, R.Bywater, G.Casari, U.Gsbel, R.W.W.Hooft, F.Kolakowski, K.Kristiansen, W.Kuipers, L.Oliveira, C.Sander, F.Rippmann, A.Valencia, E.M.van der Wenden and A.P.IJzerman for stimulating discussions. F.C. acknowledges Centre Charles Hermite and Sanofi Recherche for support. G.V. thanks the BMFT for support to the RELIWE project. We also thank Novo Nordisk Copenhagen Denmark, Solvay Weesp the Netherlands, E. Merck Darmstadt Germany and IRBM Rome Italy for testing aspects of the GPCRDB.

REFERENCES

1 Baldwin,J.M. (1993) *EMBO J.*, **12**, 1693−1703.
2 Unger,V.M. and Schertler,G.F.X. (1995) *Biophys. J.*, **68**, 1776−1786.
3 Schertler,G.F.X., Villa,C. and Henderson,R. (1993) *Nature*, **362**, 770−772.
4 Yeagle,P.L., Alderfer,J.L., Salloum,A.C., Ali,L. and Albert,A.D. (1997) *Biochemistry*, **36**, 3864−3869.
5 Yeagle,P.L., Alderfer,J.L. and Albert,A.D. (1996) *Mol. Vis.*, **2**, 12.
6 Yeagle,P.L., Alderfer,J.L. and Albert,A.D. (1995) *Biochemistry*, **34**, 14621−14625.
7 Drews,J. (1996) *Nature Biotechnol.*, **14**, 1516–1518.
8 Singer,M.S., Oliveira,L., Vriend,G. and Shepherd,G.M. (1995) *Receptors Channels*, **3**, 89−95.
9 Kuipers,W., Oliveira,L., Paiva,A.C.M., Rippman,F., Sander,C. and IJzerman,A.P. (1996) In Findlay,J. (ed.), *Membrane Protein Models*. Bios Scientific Publishers Ltd, Oxford, pp. 27−45.
10 Oliveira,L., Paiva,A.C.M. and Vriend,G. (1995) In Kazonaya,P.T.P. and Hodges,R.S. (eds), *Peptides: Chemistry, Structure and Biology.* Mayflower Scientific Ltd, Kingswinford, UK. pp. 408–409.
11 Oliveira,L., Paiva,A.C.M. and Vriend,G. (1993) *J. Comp.-Aid. Mol. Des.*, **7**, 649−658.
12 Bairoch,A. and Apweiler,R. (1997) *Nucleic Acids Res*., **25**, 31−36. [see also this issue (1998) *Nucleic Acids Res*. **26**, 38–42].
13 Kristiansen,K., Dahl,S.G. and Edvardsen,Ø. (1996) *Proteins: Struct. Funct. Genet.*, **26**, 81–94.
14 See http://www-grap.fagmed.uit.no/GRAP/homepage.html
15 See http://www.lctn.u-nancy.fr/viseur/viseur.html
16 Seeman,P. (1993) *Receptor Tables, vol.2: Drug Dissociation Constants for Neuroreceptors and Transporters.* SZ Research, Toronto.
17 Vriend,G. (1990) *J. Mol. Graph.*, **8**, 52−56.
18 Sander,C. and Schneider,R. (1991) *Proteins*, **9**, 56−68.
19 Devereux,J. (1989) *The GCG Sequence Analysis Software Package, Version 6.0*. Genetics Computer Group, University of Wisconsin Biotechnology Center, 1710 University Avenue, Madison, Wisconsin, USA, 53705.
20 Kolakowski,L.F. Jr (1994) *Receptors Channels*, **2**, 1−7.
21 Sayle,R. and Milner White,E.J. (1995) *Trends Biochem. Sci.*, **20**, 374−376.
22 See http://www4.ncbi.nlm.nih.gov/PubMed
23 On-line Mendelian Inheritance in Man (OMIM), a catalog of human gemes and genetic disorders. McKusick,V.A. *et al.* Johns Hopskins University. See http://www3.ncbi.nlm.nih.gov/omim
24 Stoesser,G., Sterk,P., Tuli,M.A., Stoehr,P.J. and Cameron,G.N. (1997) *Nucleic Acids Res*., **25**, 7–13 [see also this issue (1998) *Nucleic Acids Res*. **26**, 8–15].
25 George,D.G., Dodson,R.J., Garavelli,J.S., Haft,D.H., Hunt,L.T., Marzec,C.R., Orcutt,B.C., Sidman,K.E., Srinivasarao,G.Y., Yeh,L.S.L., *et al.* (1997) *Nucleic Acids Res*., **25**, 24−27 [see also this issue (1998) *Nucleic Acids Res*. **26**, 27–32].
26 Bairoch,A., Bucher,P. and Hofmann,K. (1997) *Nucleic Acids Res*., **25**, 217−221.
27 The FlyBase consortium (1997) *Nucleic Acids Res*., **25**, 63−66 [see also this issue (1998) *Nucleic Acids Res*. **26**, 85–88].