Chapter 1
The reincarnation of the tGRAP database

Bas Vroling

Abstract The searchable mutant database tGRAP (previously called tinyGRAP) was hosted at the University of Tromso since 1996, but went offline a few years ago. It contained a wealth of data on mutated G-protein coupled receptors (GPCRs). All data was extracted from scientific papers and entered into the database manually. In this paper, we show how webservices were used to clean and, especially, update the data available on the mutated proteins and literature references. The updated data was entered in both a relational database system and a full-text indexing system, and made publicly available on the web.

1.1 Introduction

The tGRAP [1] database contained mutant data on five families of GPRCs. Apart from information about the actual mutations in GPCRs that have been functionally tested in cellular experiments, the tGRAP database also contained qualitative information about how the mutated receptor was tested, including information about the experimental conditions. The last update (release 10) to the tGRAP database was applied in 2001. With that update, the total number of mutations available was brought to 10,500.

The database was based on flatfile entries, and queries on the entries were done through a text indexing system. An example entry file is shown below. Each mutant entry contains data on the depositor of the mutation, the receptor that was mutated, literature information and information about the mutations and experimental details and conditions.

Bas Vroling
Center for Molecular and Biomolecular Informatics e-mail: bvroling@cmbi.ru.nl

```
AU   Margot Beukers
EA   beukers@chem.leidenuniv.nl
AF   Farmacochemie, LACDR, Leiden University
AC   P30542/SwissProt
DE   ADENOSINE A1
NR   FamilyA; Nucleotidelike; Adenosine; Adenosine type1
OS   Homo sapiens (Human)
ID   AA1R_HUMAN
TO   102775
RA   Scholl DJ
RA   Wells JN
RT   Serine and alanine mutagenesis of the nine native cysteine
RT   residues of the human A(1) adenosine receptor.
RL   Biochem Pharmacol 2000; 60:1647-54
RX   11077047
SB   Cys85Ser @ TM3
SB   Cys131Ser @ TM4
SB   Cys255Ala @ TM6
SB   Cys260Ala @ XCL3
SB   Cys263Ser @ XCL3
LB   AN
OD   IM
ES   TR
AS   ME
CT   COS-M6
VE   pCMV4
MI   32051-11077047
MT   Multiple/Substitution
DR   1000
FN   ../rel10/MB1527s.txt
pr   P30542
```

As is the case with any text-based databank of reasonable size where data is entered manually, errors are bound to crawl in. In the case of the tGRAP data, it was often seen that one protein had multiple SwissProt accession codes associated with it. Another important issue, which is not a real error but a disadvantage that is frequently encountered when using this type of data entry, is the inability of the system to deal with changing information. For example, SwissProt identifiers can change over time. So, when data is not automatically updated, old and outdated information remains in the database. When these types of errors and inconsistencies accumulate, the process of retrieving useful information becomes more and more difficult.

The tGRAP database was a useful tool for planning mutant experiments, for interpretation of information from such experiments and for receptor modeling. Because of the usefulness of this large amount of data, reviving the database will be an important contribution to the GPCR research field.

1.2 Materials & Methods

In order to insure the highest possible data quality and integrity in the new database systems, the information on the mutated receptors and the literature references was updated using webservices. The Taverna [3] package was used to query the SOAP interface of the MRS [2] server for retrieving the protein information, and the PubMed SOAP interface was queried to retrieve the information on the scientific articles. Furthermore, the number of different descriptions of experimental details and conditions, such as cell types, expression systems and second messengers was greatly reduced by standardizing the nomenclature, thereby improving the consistency of the data and improving the quality of the results.

1.2.1 Accessing remote data with Taverna

The information associated with the SwissProt protein accession codes was retrieved using MRS. To implement the this update routine, the Taverna package was used. Taverna allows bioinformaticians to construct workflows or pipelines of services to perform a range of different analyses. Although the primary goal of Taverna is to ease the integration of molecular biology tools and database to answer complex questions, it can also be used to implement the relatively simple operation of updating protein information in an existing database. The workflow that was used to retrieve the data is shown in figure 1.1.

Another workflow was built to retrieve details about scientific papers from the PubMed repository. The workflow is shown in figure 1.2.

1.3 Results and discussion

The update of the protein information resulted in the retrieval of information on 260 GPCR proteins, originating from 26 species. Information on 1369 scientific papers was retrieved. The updated information was integrated in both the MRS full-text indexing system and the new, relational database driven tGRAP website. In the near future, the new tGRAP database can be easily integrated in other biological research tools and workflows, since SOAP-based webservice access for this database is in development. When complete, automation-friendly access to this important biological data will hopefully accelerate the developments in the GPCR research field, and

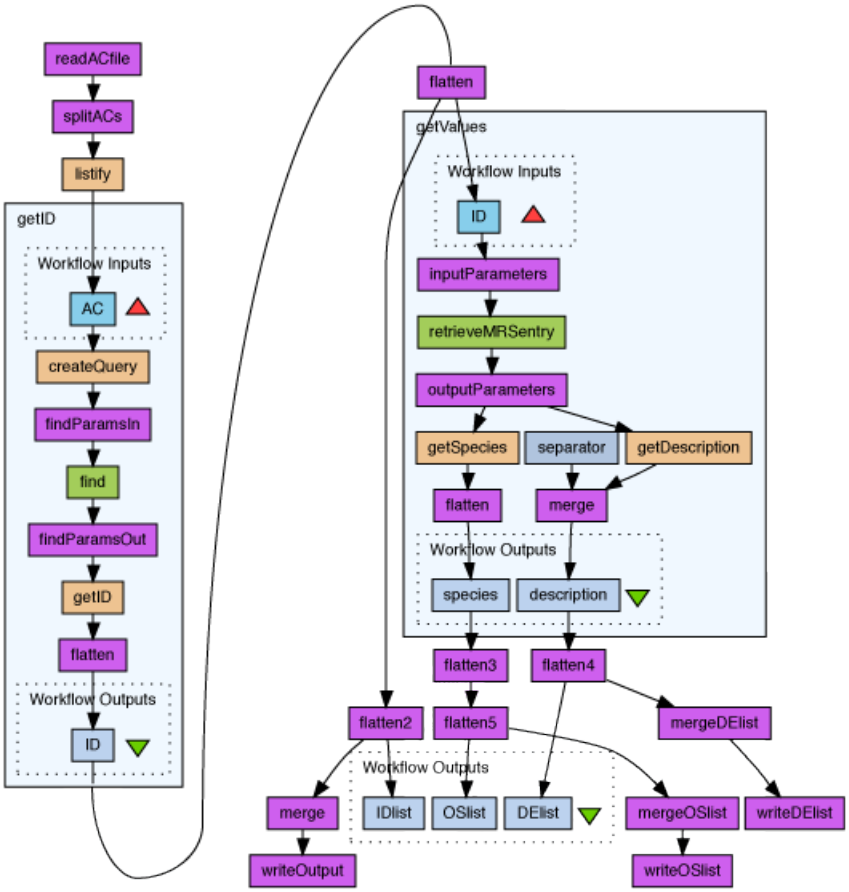

Fig. 1.1 The workflow used to retrieve updated information from SwissProt. First, MRS is queried to retrieve the corresponding IDs to the given ACs. Next, the entries are retrieved from MRS and the desired data is extracted. The 'boxed' parts are separtely stored flows, and can be reused in other flows, allowing for very modular design practice.

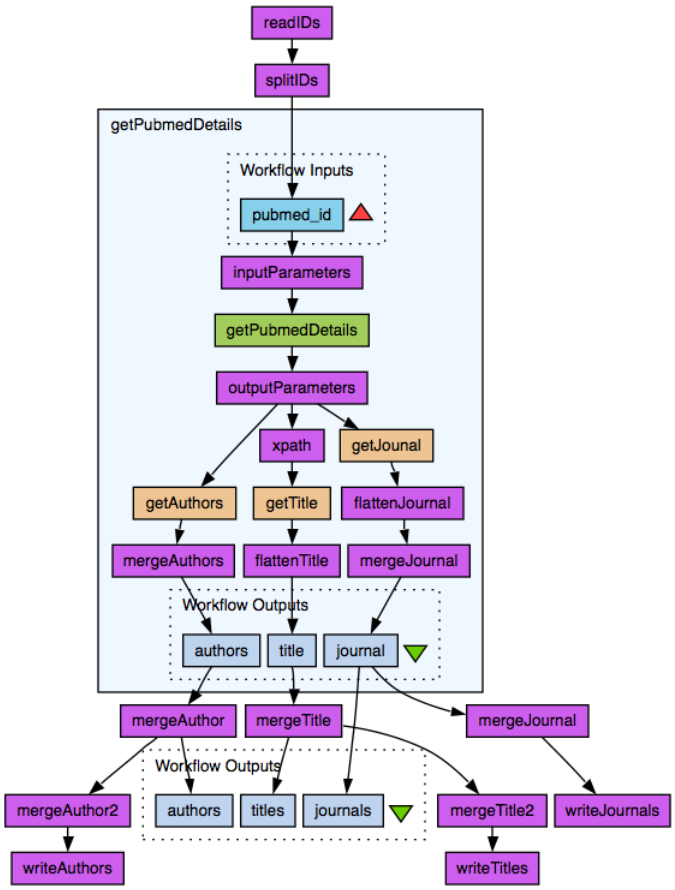

Fig. 1.2 The workflow used to retrieve scientific paper information. The PubMed IDs from the tGRAP database where used to query the PubMed webservice. From the query response, the details of the scientific paper are extracted and saved to disk.

contribute to an increased awareness of the importance of webservices for integrated bioinformatics research efforts.

References

1. Oyvind Edvardsen, Anne Lise Reiersen, Margot W Beukers, and Kurt Kristiansen. tgrap, the g-protein coupled receptors mutant database. *Nucleic Acids Res*, 30(1):361–3, Jan 2002.
2. M L Hekkelman and G Vriend. Mrs: a fast and compact retrieval system for biological data. *Nucleic Acids Res*, 33(Web Server issue):W766–9, Jul 2005.
3. Duncan Hull, Katy Wolstencroft, Robert Stevens, Carole Goble, Mathew R Pocock, Peter Li, and Tom Oinn. Taverna: a tool for building and running workflows of services. *Nucleic Acids*

Res, 34(Web Server issue):W729–32, Jul 2006.