THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Retrieving GPCR data from public databases

Retrieving GPCR data from public databases

Christopher Southan

IUPHAR/BPS Guide to PHARMACOLOGY, Centre for Integrative Physiology, University of Edinburgh, EH8 9XD, UK

Corresponding author: Dr C Southan cdsouthan@hotmail.com

Abstract

Improvements in databases have already impacted GPCR research. The purpose of the review is to give a snapshot of the GPCR data available and provide utility examples Consequently, this review covers a small set of major databases, including UniProt for proteins, Ensembl for genes, ChEMBL for bioactive chemistry and SureChEMBL for patents. In addition, two portals are outlined, GPCRdb and the IUPHAR/BPS Guide to PHARMACOLOGY (GtoPdb) that are based on expert annotation. The former has an emphasis on structures, sequences, point mutations, analysis tools and visualisation. The latter focuses on endogenous GPCR ligands, pharmacological modulation, approved drugs clinical candidates and tool compounds. Since data growth is accelerating, those embarking on GPCR projects should not only check databases but also recent journal and patent publications.

Introduction

Over the last few decades, databases have moved from the periphery of the biomedical sciences to a central position [1]. This digitally-driven migration is in part due to data deposition requirements of journals, improvements in content, usability, access, documentation and automated interoperability. This progress is having an increasing impact on both experimental and *in silico* research [2]. GPCR-related databases have been previously reported in detail up to 2013 [3]. This article focuses on a key set of open resources that have updated in the last three years (Fable 1). Some of these have introductory and educational value but this is outside of the scope of this review. It should also be noted that GPCR researchers may have divergent foci. For example, these could encompass receptor evolution, biochemical mechanisms (including de-orphanisation), disease involvement, drug discovery, chemical biology and clinical pharmacology. A classification that can make matching to particular research interests easier is to divide databases into primary, secondary and tertiary (analogous to journal research reports, review articles and book chapters). While this division is not strict, it is useful to indicate differences in scales of data collation as well as the balance between manual curation versus automated annotation. The selected sources are listed in Table 1. The Endothelin A receptor (ET$_A$ receptor or ENDRA, see other synonyms in Table 1) will be used as a human GPCR example since its associated database records have recently been reviewed [4].

Primary Databases

Primary databases are associated with the direct deposition of experimental data at large scale. They rely heavily on automated annotation and are updated frequently. Their value lies in maximal coverage but they can be challenging to navigate. The best known example is the GenBank repository for nucleic acid sequences [5]. This has 314 million entries from which "endothelin receptor" will retrieve 1518. One of these is a 69Kb genomic DNA sequence (GenBank: AY422989) that includes annotation of the ENDRA gene structure. We can also find S63938, an 1868 base pair placental mRNA, as one of the transcript depositions for ENDRA. The protein translation, also a GeneBank record, is an open reading frame (ORF) of 427 amino acids designated AAB20278. Within the coding exons we can find a single nucleotide polymorphism (SNP) designated as rs772147672 from the 145 million reference

entries in the dbSNP database [6]. It should be pointed out that 3D structures are also primary data. Since there is no ENDRA crystal structure (yet), we can use the recent example of the M4 muscarinic acetylcholine receptor bound to tiotropium, released as PDB: 5DSG in March 2016 [7]. It should also be noted that PubChem can be considered a primary chemistry resource with over 90 million structures [8]. Different query routes can be used to find bioactive small-molecules acting on ENDRA but the Entrez Gene identifier 1909 links to 48 PubChem Compound Identifiers (CIDs). Following the molecular details for GPCRs through primary databases, as illustrated in this paragraph, provides comprehensive and up to date coverage. However, since this requires extensive bioinformatics expertise, users may prefer the integrated presentation of this data in secondary databases.

Secondary Databases

Secondary databases are value-added aggregations of primary data sources that use automated merging rules together with some manual curation. They are typically orders of magnitude smaller than primary databases and consequently easier to navigate. Examples for proteins, genes, chemistry, patents and diseases are selected here, as judged on their reliability. The first is UniProtKB containing over 65 million proteins extracted from GenBank. The quality of the ~0.5 million manually reviewed entries that constitute the Swiss-Prot section, makes UniProt the first choice of GPCR secondary protein resources [9]. It massively reduces the redundancy in primary sources (e.g. multiple mRNAs, alternative splicing, polymorphisms, PDB structures, pathways and bioactive chemistry) since these are all merged as annotated features or links from a single protein entry. For the human ENDRA entry (P25101, EDNRA_HUMAN), 22 GenBank primary mRNA and genomic DNA entries are cross-referenced. UniProt coverage can be assessed with the keyword "g protein-coupled receptor". This returns 101,227 non-reviewed (i.e. automatically annotated) and 3,250 manually reviewed (Swiss-Prot) entries, of which 830 are human. Adding the search restriction (NOT) "olfactory receptor" reduces this to 403.

The chosen genome resource is the EBI/Sanger Centre Ensembl where ENDRA can be selected via the identifier ENSG00000151617 [10]. Navigation (via zooming and track-toggling) can be used to discern aspects of gene evolution, regulation, expression, splicing and variation. Note that (with quality filtration) 23,277 dbSNP entries are mapped onto the ENDRA gene locus. New versions of Ensembl, including gene (re)builds synchronised with

new genome assemblies and variation data, are released every three months. Given the relative stability of the human reference genome, GPCRs are likely to have a consistent genetic architecture (i.e. the same spacing between introns, exon and regulatory regions) between releases, However, releases can have minor coordinate changes as well as new annotations for splicing, polymorphisms, improved resolution of regulatory regions, disease associations and orthologue expansions (via more species).

ChEMBL can be considered a secondary database with a focus on collating structure-activity relationship (SAR) data from the medicinal chemistry literature. Release 21 has extracted not only 1.1 million chemical structures and associated results from just over 62,000 papers but also ~0.5 million imported from confirmatory PubChem BioAssays [11]. We can establish that 2,577 compounds are mapped to the target entry for ENDRA (CHEMBL252) and that 231 (non-olfactory) human GPCRs have a ChEMBL link. Note that ChEMBL submits structures and BioAssay data to PubChem within a few weeks of a new release. Given the importance of chemistry-to-GPCR mapping (as indicated by the other articles in this issue) we can also count mappings for three other bioactivity databases with UniProt cross-references (Fig. 1).

The DrugBank [12] and BindingDB [13] resources are described in recent publications. Differences in relationship mapping between ChEMBL and DrugBank have also been detailed [14], with ChEMBL recently publishing their own quality analysis [15]. Regardless of complexities (e.g. where interactions include endogenous ligands, exogenous peptides, clinical antibodies and some inactive results) Figure 1 indicates that ~2/3rds of the 403 non-olfactory GPCRs have some kind of activity modulators.

In terms of bioactive chemistry, GPCR researchers may neither be aware that the patent corpus documents at least twice as much SAR as journal papers, nor that PubChem now contains over 20 million patent-extracted structures [16]. There are many options to retrieve GPCR-related patents but because the documents are linked to over 17 million structures SureChEMBL has become the secondary database of choice [17]. As an example, WO2009024906 from Actelion claims CID 25099191 as an ENRA antagonist with a 3.4 nM IC_{50} value, which was later published and extracted as CHEMBL2165326. It also happens to be a metabolite of the approved ENDRA antagonist macitentan (CID 16004692). While searching patents remains a challenge for non-specialists, back-mapping is somewhat

easier. This means that for SAR published in a journal (e.g. captured by ChEMBL or GtoPdb) the structures can often be matched, via PubChem or SureChEMBL, back to an earlier patent publication with more data. Note also that SureChEMBL updates *in situ* within a week of patent publication.

Of the database relationships GPCR researchers might want to explore, disease associations seem to be among the most challenging. Reasons include the absence of a primary sources equivalent to molecular data, the difficulty of integrating Mendelian disorders with the flood of genome-wide diseases association (GWAS) results and problems of confirming mutation-to-phenotype causality [18].   Of the many sources in this domain, ClinVar is a useful secondary database that collates relationships from multiple sources of human variations and phenotypes, together with supporting evidence [19]. For ENDRA, 17 entries (including complex structural genomic variants) can be accessed via the link from the [Entrez Gene ID 1909](#) . However, the SwissProt ENDRA entry points just to three "Natural variant" amino acids exchanges, one of which is somatic rather than germline [20] . Given the acceleration in genome sequencing of all types of disease cohorts (including for rare diseases) an increase in GPCR variants with clinical effects is expected to be captured in ClinVar, and Swiss-Prot.

Tertiary Databases

While there is no clear division between secondary and tertiary sources, the latter integrate the former with a thematic focus. They are also smaller scale, include a higher density of expert curation and - commensurate with their utility as first-stop portals - have a strong focus on user navigation. Two of these will be outlined here, GPCRdb [21] [22] and the IUPHAR/BPS Guide to PHARMACOLOGY (GtoPdb) [23]. The selection criteria are a) they are well established, b) have detailed 2016 papers c) undergo frequent updates, d) directly collaborate on complementary utility, including web services compatibility e) they extensively and reciprocally cross-point to other resource f) are well documented and g) freely accessible.

 The range of GPCRdb features includes the following;
1. An emphasis on sequences, structural information and analysis tools
2. Manual curation of the core annotation integrated with computationally-derived data

3. Alignments with ~ 18,000 orthologues from UniProt updated when new PDB structures become available

4. These are used as templates to optimise alignments and standardise generic residue numbering (142 are loaded for 37 unique GPCRs so far) [24]

5. In addition to literature extraction of experimental mutation data, users can directly submit standardised result sets.

6. Displays to support mutagenesis work include snake-plots of mutant proximity to ligand binding sites and tables of positions across paralogues and orthologues.

7. Additional representations assess ligand selectivity and important interactions

8. Pharmacophore analysis tools that use ligand fragments from PDB structures [25]

9. Links to external GPCR modelling servers

One of the key differences between GPCRdb and GtoPdb is that the latter covers all human target classes. The 2016.2 release (March) includes 14,327 curated interactions (mostly IC_{50}, K_i, and K_d measurements) across 2,775 proteins and 8,400 ligands. The following points will highlight the GPCR content.

1. The current release has 245 Human GPCR UniProt IDs with quantitative interactions for over 3,500 distinct ligands

2. 105 ligands have additional interactions to non-GPCRs (e.g. transporters)

3. GPCR-directed approved drugs, clinical candidates and tool compounds are included

4. Relationships captured by expert manual curation include free-text comments

5. Annotation is supported by target family subcommittees of the International Union of Basic and Clinical Pharmacology Committee on Receptor Nomenclature and Drug Classification (NC-IUPHAR). The 60 GPCR subcommittees involve ~540 contributing scientists

6. Selective compounds for *in vitro* and *in vivo* experimentation are highlighted

7. Inclusion of manually checked reciprocal links to key genomic, protein and small molecule resources (e.g. GPCRdb, UniProt and PubChem) as well as a "Useful Links" page

8. Monitoring de-orphanisation and/or surrogate ligands for GPCRs

9. Curation of major disease-associated clinical variants

10. Source for the biennially published "Concise Guide to PHARMACOLOGY" series of reviews, including 2015/16: G protein-coupled receptors [26]

Both these tertiary resources not only welcome technical feedback on their entries but also present additional options for user engagement. For GtoPdb, investigators can take the opportunity to join NC-IUPHAR GPCR family-specific subcommittees and/or submit their newly published papers. GPCRdb can be contacted for the direct submission of pre-publication mutagenesis data.

Conclusions

GPCR researchers have an expanding choice of databases from which limited selection is introduced here. Those less familiar with sequences, protein structures and bioactive chemistry may find the navigation order tertiary > secondary > primary easier that the other way round, particularly as the two tertiary portals GPCRdb and GToPdb described provide "first-stop-shops" from either structure-centric or pharmacology-centric viewpoints. Notwithstanding, while they offer massive efficiency gains, users also need to be aware of the shortcomings of databases. Particularly for GPCRs, it is clear that detailed review papers (including those published under the auspices of NC-IUPHAR) distil knowledge, perspectives, nomenclature details, pharmacological complexities and bioassay nuances that are difficult to distil into structured database records. Users should also appreciate that database content lags behind journal publications (e.g. the MeSH term "Receptors, G-Protein-Coupled" retrieves 7322 entries just for 2015) and secondary databases can only capture subsets of these (i.e. PubMed, PubMed Central and European PubMed Central can be considered primary databases). Consequently, where experimentation is being planned on the basis of database findings, it is prudent to check the newest primary data entries and monitor the recent literature.

Conflict of interest statement

None declared

Acknowledgments

Figures

Graphical abstract
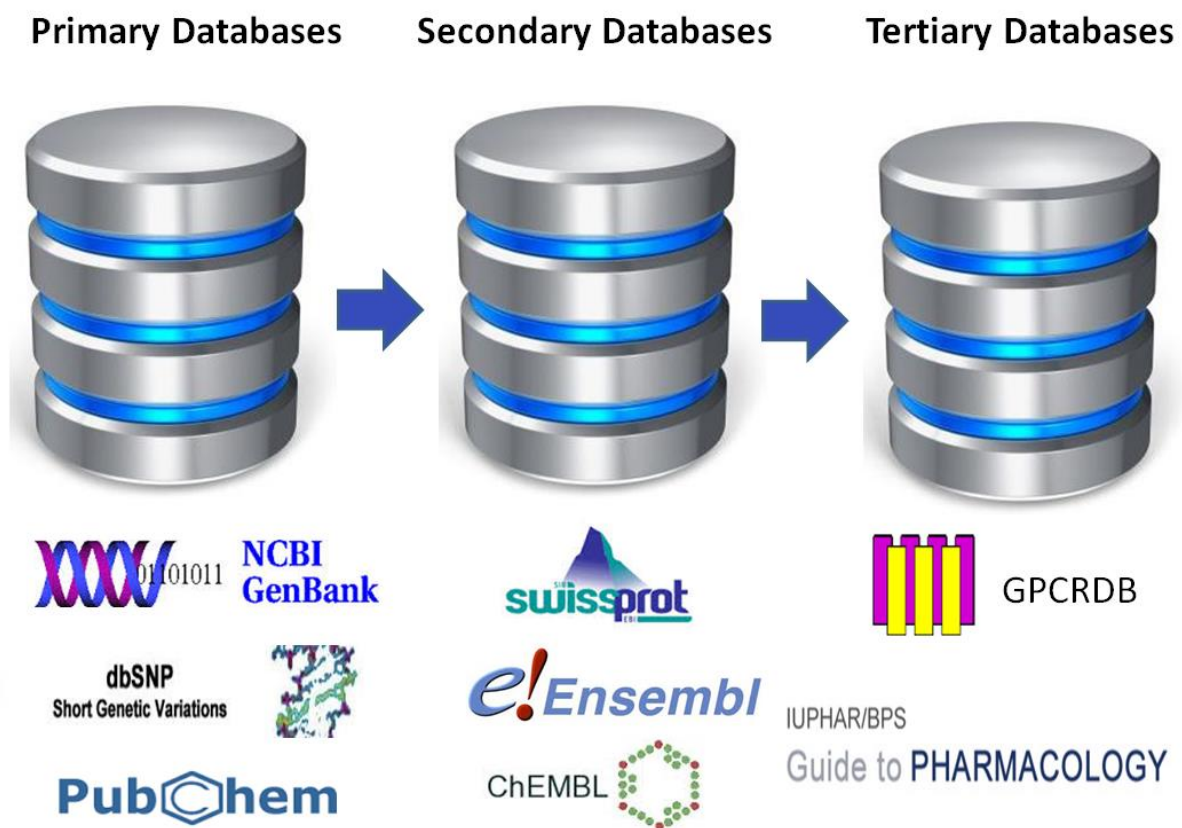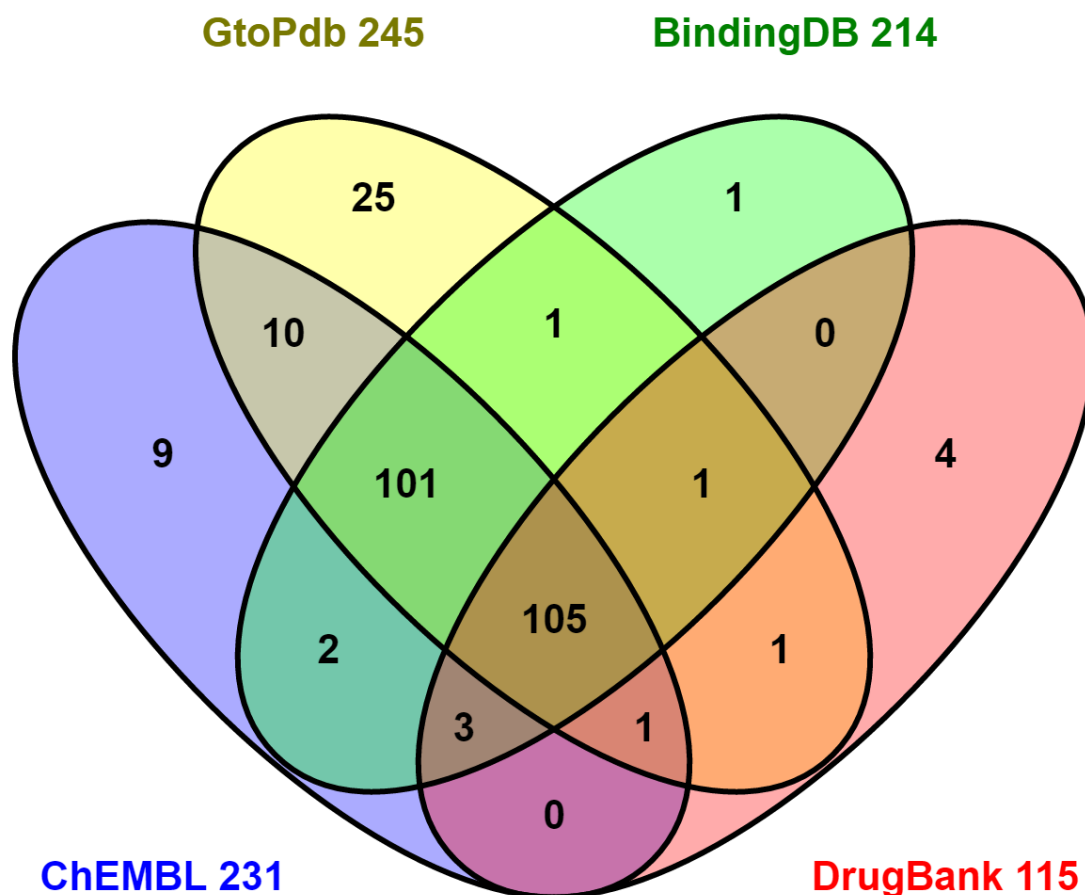
Figure 1. Comparing UniProt chemistry database cross-references for GPCRs. The Venn diagram shows the overlaps and differences (GtoPdb = IUPHAR/BPS Guide to PHARMACOLOGY). Three sources include activity data for the chemistry-to-protein interactions but DrugBank indicates only relationships. Individual totals are given in the source labels. The sum of all four extends to 265 GPCRs with a 4-way consensus intersect of 105.
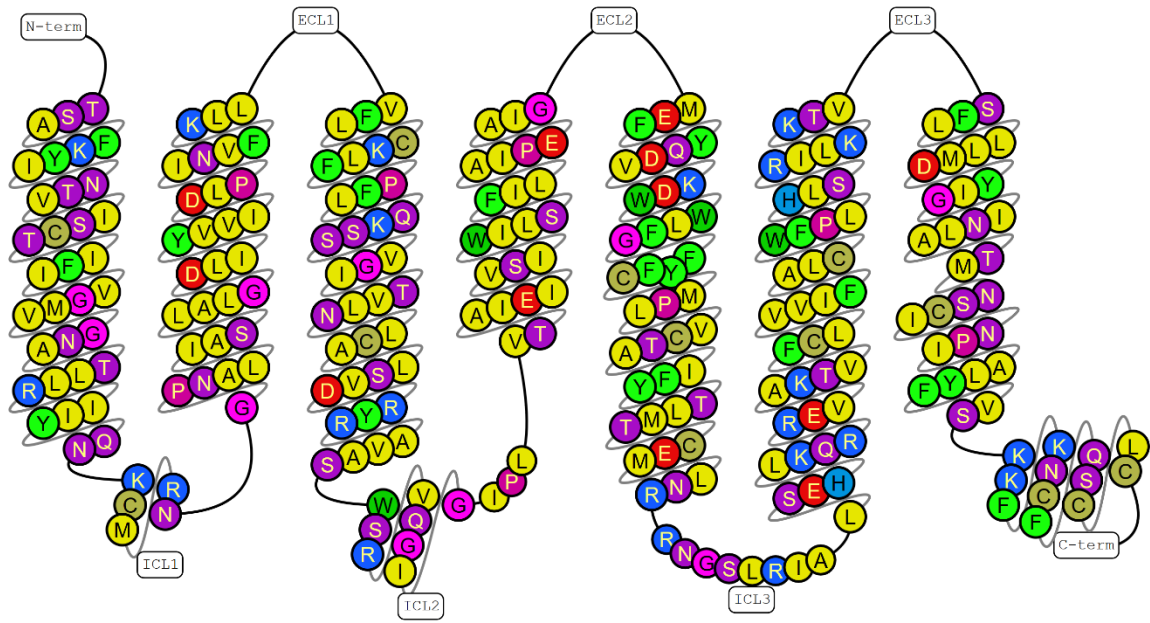
Figure 2. Screenshot from the GPCRdb ENDRA entry
(http://gpcrdb.org/protein/ednra_human/), the coloured residue properties are displayed.
Users can alternatively select 51 mutation data points to be indicated on the snake plot or
tabulated with source link-outs.

Antagonists at human ET$_A$ receptor

Key to terms and symbols   Click column headers to sort

Ligand	Sp.	Action	Affinity	Units	Reference	
atrasentan	Hs	Antagonist	9.2 – 10.5	pA_2	44	
PD-156707	Hs	Antagonist	8.1 – 9.2	pA_2	36,46	
sitaxsentan	Hs	Antagonist	8.0	pA_2	55	
BQ123	Hs	Antagonist	6.9 – 7.4	pA_2	35	
ambrisentan	Hs	Antagonist	7.1	pA_2	6	
[^{125}I]PD164333	Hs	Antagonist	9.6 – 9.8	pK_d	15	▼
PD-156707	Hs	Antagonist	9.0 – 9.8	pK_d	36	▼
[^{125}I]PD151242	Hs	Antagonist	9.0 – 9.1	pK_d	16	▼
PD164333	Hs	Antagonist	9.0	pK_d	15	▼
[^{3}H]BQ123	Hs	Antagonist	8.5	pK_d	29	▼
atrasentan	Hs	Antagonist	10.5	pK_i	44,54	▼
darusentan	Hs	Antagonist	8.9	pK_i	48	
sparsentan	Hs	Antagonist	8.0	pK_i	38	▼
macitentan	Hs	Antagonist	9.3	pIC$_{50}$	5	▼
SB234551	Hs	Antagonist	8.7 – 9.0	pIC$_{50}$	41	
zibotentan	Hs	Antagonist	8.3	pIC$_{50}$	39	▼
ambrisentan	Hs	Antagonist	7.7	pIC$_{50}$	6	▼
FR139317	Hs	Inverse agonist	7.3 – 7.9	pIC$_{50}$	35	▼
avosentan	Hs	Antagonist	7.3	pIC$_{50}$	7	▼

Figure 3. Screen shot of selected panels from the GtoPdb ENDRA entry.
http://www.guidetopharmacology.org/GRAC/ObjectDisplayForward?objectId=219
The contents panel list the curated relationships. Below this is a section on gene and protein information with 25 out-links. The lower section includes just six rows from 29 agonists and antagonists (hovering the mouse provides a key to the symbols)

11

References

1.   Southan C, Cameron G: **Beyond the Tsunami: Developing the Infrastructure to Deal with Life Sciences Data**. In *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Edited by Hey T, Tansley S, Tolle KM. Microsoft Research; 2009:117–123.

2.   Kafkas Ş, Kim J-H, McEntyre JR: **Database citation in full text biomedical articles. [Internet]**. *PLoS One* 2013, **8**:e63184.

3.   Kowalsman N, Niv MY: **GPCR & company: databases and servers for GPCRs and interacting partners.** *Adv. Exp. Med. Biol.* 2014, **796**:185–204.

4.   Davenport AP, Hyndman KA, Dhaun N, Southan C, Kohan DE, Pollock JS, Pollock DM: **Endothelin**. *Pharmacol. Rev.* 2016, **68**:357–418.

5.   Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank.** *Nucleic Acids Res.* 2015, **44**:D67–72.

6.   Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Res.* 2001, **29**:308–11.

7.   Velankar S, van Ginkel G, Alhroub Y, Battle GM, Berrisford JM, Conroy MJ, Dana JM, Gore SP, Gutmanas A, Haslam P, et al.: **PDBe: improved accessibility of macromolecular structure data from PDB and EMDB.** *Nucleic Acids Res.* 2016, **44**:D385–95.

8.   Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, et al.: **PubChem Substance and Compound databases.** *Nucleic Acids Res.* 2016, **44**:D1202–13.

9.   **UniProt: a hub for protein information** *Nucleic Acids Res.* 2014, **43**:D204–12.

10. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, et al.: **Ensembl 2016.** *Nucleic Acids Res.* 2015, **44**:D710–716.

11. Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, Krüger FA, Light Y, Mak L, McGlinchey S, et al.: **The ChEMBL bioactivity database: an update.** *Nucleic Acids Res.* 2014, **42**:D1083–90.

12. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, MacIejewski A, Arndt D,

Wilson M, Neveu V, et al.: **DrugBank 4.0: Shedding new light on drug metabolism**. *Nucleic Acids Res.* 2014, **42**.

13. Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J: **BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology.** *Nucleic Acids Res.* 2015, doi:10.1093/nar/gkv1072.

14. Southan C, Sitzmann M, Muresan S: **Comparing the Chemical Structure and Protein Content of ChEMBL, DrugBank, Human Metabolome Database and the Therapeutic Target Database**. *Mol. Inform.* 2013, **32**:881–897.

15. Papadatos G, Gaulton A, Hersey A, Overington JP: **Activity, assay and target data curation and quality in the ChEMBL database**. *J. Comput. Aided. Mol. Des.* 2015, **29**:885–896.

16. Southan C: **Expanding opportunities for mining bioactive chemistry from patents**. *Drug Discov. Today Technol.* 2015, **14**:3–9.

17. Papadatos G, Davies M, Dedman N, Chambers J, Gaulton A, Siddle J, Koks R, Irvine SA, Pettersson J, Goncharoff N, et al.: **SureChEMBL: a large-scale, chemically annotated patent document database.** *Nucleic Acids Res.* 2015, doi:10.1093/nar/gkv1253.

18. Vihinen M, Hancock JM, Maglott DR, Landrum MJ, Schaafsma GCP, Taschner P: **Human Variome Project Quality Assessment Criteria for Variation Databases.** *Hum. Mutat.* 2016, **37**:549–58.

19. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J, et al.: **ClinVar: public archive of interpretations of clinically relevant variants.** *Nucleic Acids Res.* 2016, **44**:D862–8.

20. Mottaz A, David FPA, Veuthey AL, Yip YL: **Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar**. *Bioinformatics* 2010, **26**:851–852.

21. Munk C, Isberg V, Mordalski S, Harpsøe K, Rataj K, Hauser A, Kolb P, Bojarski AJ, Vriend G, Gloriam DE: **GPCRdb: The G protein-coupled receptor database - An introduction.** *Br. J. Pharmacol.* 2016, doi:10.1111/bph.13509.

22. Isberg V, Mordalski S, Munk C, Rataj K, Harpsøe K, Hauser AS, Vroling B, Bojarski AJ, Vriend G, Gloriam DE: **GPCRdb: an information system for G protein-coupled receptors.** *Nucleic Acids Res.* 2016, **44**:D356–64.

23. Southan C, Sharman JL, Benson HE, Faccenda E, Pawson AJ, Alexander SP, Buneman OP, Davenport AP, McGrath JC, Peters JA, et al.: **The IUPHAR/BPS**

Guide to PHARMACOLOGY in 2016: towards curated quantitative interactions between 1300 protein targets and 6000 ligands *Nucleic Acids Res* 2016, **44**:D1054–68.

24.   Isberg V, De Graaf C, Bortolato A, Cherezov V, Katritch V, Marshall FH, Mordalski S, Pin JP, Stevens RC, Vriend G, et al.: **Generic GPCR residue numbers - Aligning topology maps while minding the gaps**. *Trends Pharmacol. Sci.* 2015, **36**:22–31.

25.   Fidom K, Isberg V, Hauser AS, Mordalski S, Lehto T, Bojarski AJ, Gloriam DE: **A new crystal structure fragment-based pharmacophore method for G protein-coupled receptors**. *Methods* 2015, **71**:104–112.

26.   Alexander S, Sharman JL, Spedding M, Peters JA, Harmar AJ: **the Concise Guide To Pharmacology 2013 / 14 : GPCRs**. *Br. Jourmal Pharmacol.* 2013, doi:10.1111/bph.12444/full.

Recommended reading – authors' annotations

*Of special interest

Kowalsman N et al (3) Extensive (but non-selective) review of 40 GPCR-related resouces

Davenport et al (4) Detailed review of endothelin, its receptors and their pharmacology.

UniProt Consortium. (9) Updates and new features.

Yates A, et al (10) Includes new features for comparing GPCRs across species.

Southan C et al (14) A detailed study of four databases with GPCR-related content

Southan C (16) Outlining the "big bang" of patent chemistry in PubChem.

Papadatos G et al (17) Desciption of a leading patent extraction resource.

Landrum MJ et al   (19) Expanding resource of clinically relevant genetic variants

**Of outstanding interest

Kim S, et al (15) Largest primary source of small molecules and peptides acting on GPCRs,

Bento et al (11) Describes large-scale extraction of medicinal chemistry data from papers.

Isberg V et al (22) Report on new features, displays and tools.

Southan C et al (23) Content capture, relationship statistics and curation rules for GtoPdb

Alexander SPH et al (26) A review snapshot of GPCR entries extracted from GtoPdb.