



John Overington

*PhD in Crystallography, Birkbeck College, London, 1991.
Postdoctoral research, ICRF, 1990-1992. Pfizer 1992-2000.
Inpharmatica 2000-2008.
At EMBL-EBI since 2008.*

ChEMBL

DESCRIPTION OF SERVICES

The ChEMBL group develops and manages the EBI's database of bioactive, drug-like small molecules, which contains two-dimensional structures, calculated properties and abstracted bioactivities such as binding constants, pharmacology and ADMET data. ChEMBL data are abstracted and curated from the primary scientific literature, and cover a significant fraction of the structure-activity relationship and discovery of modern drugs. 2010 marked the first full year of staffing for the group, and has seen a number of milestones for the ChEMBL resource. Its first public release – in January 2010 – achieved broad coverage in the press and was extremely well received by the scientific community. The data is widely accessed via the web interface and via download of the entire database for local searching, and advanced tools developed by the ChEMBL team for interactive filtering and data selection provide added value to users.

SUMMARY OF PROGRESS

- Switched all ChEMBL resources to run under the secure, industry standard https: internet protocol;
- Established a robust, monthly update cycle for new data;
- Established a mechanism for the rapid upload, archival and searching of deposited datasets;
- Launched SARfari drug-discovery integration systems;
- Achieved integration of ChEMBL into other large-scale chemistry resources, including PubChem and the ChemSpider system of the Royal Society of Chemistry;
- Started to implement a fully featured and open infrastructure for large-scale scoring of targets for their 'drugability';
- Pursued research activities in two major areas.

MAJOR ACHIEVEMENTS

Usage of the resource, in particular downloads of the data, has been strong and steady. We accompanied the ChEMBL launch with a series of talks, webinars, on-campus training courses and site visits for local training. Alongside more traditional approaches to promoting resource awareness, we maintain a group blog to report on progress with the database, new drug launches and various analyses of the resource.

We switched all ChEMBL resources to run under the secure, industry standard https: internet protocol, ensuring that all traffic to our services is encrypted and secure. This is a key concern for researchers given the high confidentiality of small molecule structural data. We also established a robust monthly update cycle for new data, giving the community rapid access to new chemotype and target information. A network of specialist curators has been engaged to curate key portions of the data (e.g. ADMET data). During the reporting period the number of data records within ChEMBL grew by more than 50%.

The group established a mechanism for the rapid upload, archival and searching of deposited datasets. This year, three deposited datasets on whole-cell malaria screening – contributed by GlaxoSmithKline, Novartis and St. Judes – featured ca. 20 000 novel compounds that are active in a relevant model of malaria infection. A specific portal, ChEMBL-NTD (Neglected Tropical Diseases), was constructed to serve this important subset of contributed data.

We went live with our SARfari drug discovery integration systems, two of which integrate data for bioassays, phylogenetic information, three-dimensional structural data and binding-site data for protein kinases and rhodopsin-like G-protein coupled receptors (GPCRs). In addition, the ChEMBL database began to be widely integrated into other large-scale chemistry resources. Of particular note is the integration of its chemical structure and bioactivity data into PubChem, as well as compound-level integration with the ChemSpider system of the Royal Society of Chemistry.

One of the areas of most immediate application for the data contained within ChEMBL is in the assessment and scoring of proteins as targets for drug discovery. We have started to implement a fully featured and open infrastructure for large-scale scoring of targets for their 'drugability'. The first released component is an analysis of properties of the binding sites for their suitability to bind drug-like molecules.

We currently have two active research areas. The first is the building of a computational system to analyse functional and binding data for peptides, and then to propose their optimisation in order to improve pharmaceutical properties, stability, affinity and selectivity. We published a paper on the analysis of ligand efficiency measures for the content of ChEMBL as well as a series of similarity maps for natural and unnatural amino acids. This work is funded under the EIPOD scheme, with the designed peptides planned for synthesis and bioassay in the lab of Maja Koehn (EMBL-Heidelberg).

The second area of research is a comprehensive analysis of 'tool compounds' or 'chemical probes'. We have assembled a number of sets of compounds that are generally considered to be chemical tools, that is, small molecules that are used to probe the function of specific proteins in either a cell or an in vivo model system. These compounds have been characterised for various properties (e.g. affinity, molecular size); approaches to predict the affinity variances across model organism species have been developed (i.e. across rat, mouse, and human orthologues).

We have participated in two significant EU-funded projects: eTox and EU-OPENSREEN. eTox is an Innovative Medicines Initiative that aims to build an unprecedented collaborative database of chronic rat-toxicity data and then perform bio- and chemoinformatic analyses and software development to predict toxicity, thereby improving the productivity of pharmaceutical discovery. EU-OPENSREEN is a large-scale infrastructure; EMBL-EBI is involved defining its data-standards (see the Steinbeck group, page 54) and database design and content (ChEMBL group). This project will provide an open-access and open-data infrastructure for screening a large compound collection and disseminating the collected data.

FUTURE PLANS

This coming year, we will release the drugability prioritisation and analysis tools, and also populate the database with biotherapeutic and clinical candidate development data. Also of high priority will be completing integration with core EMBL-EBI resources such as Ensembl, UniProt, PDB and ArrayExpress.

SELECTED REFERENCES

Abad-Zapatero, C., et al. (2010) Ligand efficiency indices for an effective mapping of chemo-biological space: the concept of an atlas-like representation. *Drug Discov. Today* 15, 804-811.

Gaulton, A. and Overington, J.P. (2010) Role of open chemical data in aiding drug discovery and design. *Future Med. Chem.* 2, 903-907.

ChEMBL NTD Compound Search Results: 13519 Hits

Mini Report Cards << 1 >> Please select...

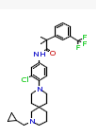
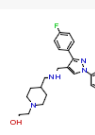
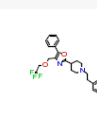
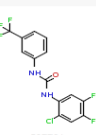
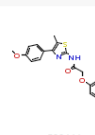
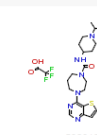
Compound	Sources	Synonyms	Mol Weight	ALogP	PSA	#Ro5 Vio.	%iHB 3D7(2uM)	%iHB Dd2(2uM)	%iHB 3D7 PFLDH (2uM)	pXC50 3D7	%iHB HEPG2(10uM)	IFI	Chemical Cluster NR	Graph Frame Cluster	Annotations
 525301	TCMDC-134639	625301 (C...)	548.08	7.12	35.58	3	0	4	0	5.69 (1999.08nM)	0	6.67	2143	239	(Predicted) Ion channel inhibitor
 542704	TCMDC-124688	642704 (C...)	422.54	4.38	53.32	4	0	0	0	5.69 (1998.48nM)	29	11.36	764	258	
 541334	TCMDC-131904	641334 (C...)	581.80	6.34	67.60	4	0	3	0	5.70 (1969.06nM)	9	0.72	381	293	
 537761	TCMDC-123779	637761 (C...)	350.67	4.54	41.13	1	2	0	0	5.70 (1958.71nM)	0	11.56	153	369	
 530444	TCMDC-124330	630444 (C...)	384.45	3.51	97.92	5	1	0	0	5.71 (1948.23nM)	0	1.85	138	248	
 532013	TCMDC-134390	632013 (C...)	513.10	4.83	92.84	5	1	7	0	5.71 (1943.48nM)	6	5.66	208	236	

Figure. Representative screen-shot of the ChEMBL database showing flexible querying and powerful analysis routines for bioactivity data.