ELSEVIER

editorial

Ye Hu

Jürgen Bajorath

# Learning from 'big data': compounds and targets

The 'big data' wave, for several years already an intensely discussed topic in bioinformatics [1], is hitting pharmaceutical R&D [2]. Exponential growth of data volumes in the life sciences and the informatics challenges that come with handling these data are no news. However, there is more to the big data deluge than mere volumes; in particular, increasing data heterogeneity and complexity makes it difficult to extract knowledge from such data. If the use of big data for drug discovery should indeed open new frontiers, and not only be hype, new visions and concepts are required to reduce data complexity and increase data consistency from different sources. Let us consider an example. It has recently been shown that two independent large-scale cancer cell line profiling studies sharing subsets of cell lines, drugs and gene expression data displayed a high degree of correlation between gene expression profiles (although different microarray platforms were used), but only little correlation between (related yet distinct) drug sensitivity assays [3], leading to rather different gene–drug associations. With increasing data complexity, such discrepancies can – and should – be expected, even in the presence of high experimental/technical quality.

Compound activity data principally have lower inherent complexity than, for example, gene–drug associations, but must also be considered in the context of big data relevant for pharmaceutical R&D. For example, the current release 17 of the ChEMBL database [4] contains >1.3 million unique compounds having >12 million activity annotations for ~9300 biological targets. Moreover, in PubChem's Compound [5], Substance [5], and BioAssay [6] databases, there currently are ~49 million compound, ~128 million substance, and ~740,000 assay entries, respectively (with daily increasing counts). The PubChem assays are associated with ~103,000 target proteins (with different sequences). In addition, there are 3846 confirmatory bioassays involving 2533 different targets. These numbers alone, without taking any other compound data sources into consideration, demonstrate the advent of 'big compound data'. In addition to large volumes, compound activity data involving different types of activity measurements, assays of varying activity and/or target confidence, and differently defined target annotations are also heterogeneous and complex, which is often not sufficiently considered when retrieving activity information from compound repositories.

As an example, let us have a look at publicly available activity/target annotations for exemplary approved drugs, which are, by definition, among the best characterized small molecules. Here, DrugBank [7] is utilized as a primary source for these drugs and, in addition to ChEMBL and PubChem, activity data is also obtained from BindingDB [8] and Open PHACTS [9]. A pharmacology record in Open PHACTS reports a biological target and/or activity for a given compound. In Fig. 1a, a pair of structurally analogous drugs is shown (trimeprazine and promethazine, both of which are applied for the treatment of allergic disorders). In addition,

1359-6446/06/$ - see front matter  http://dx.doi.org/10.1016/j.drudis.2014.02.004

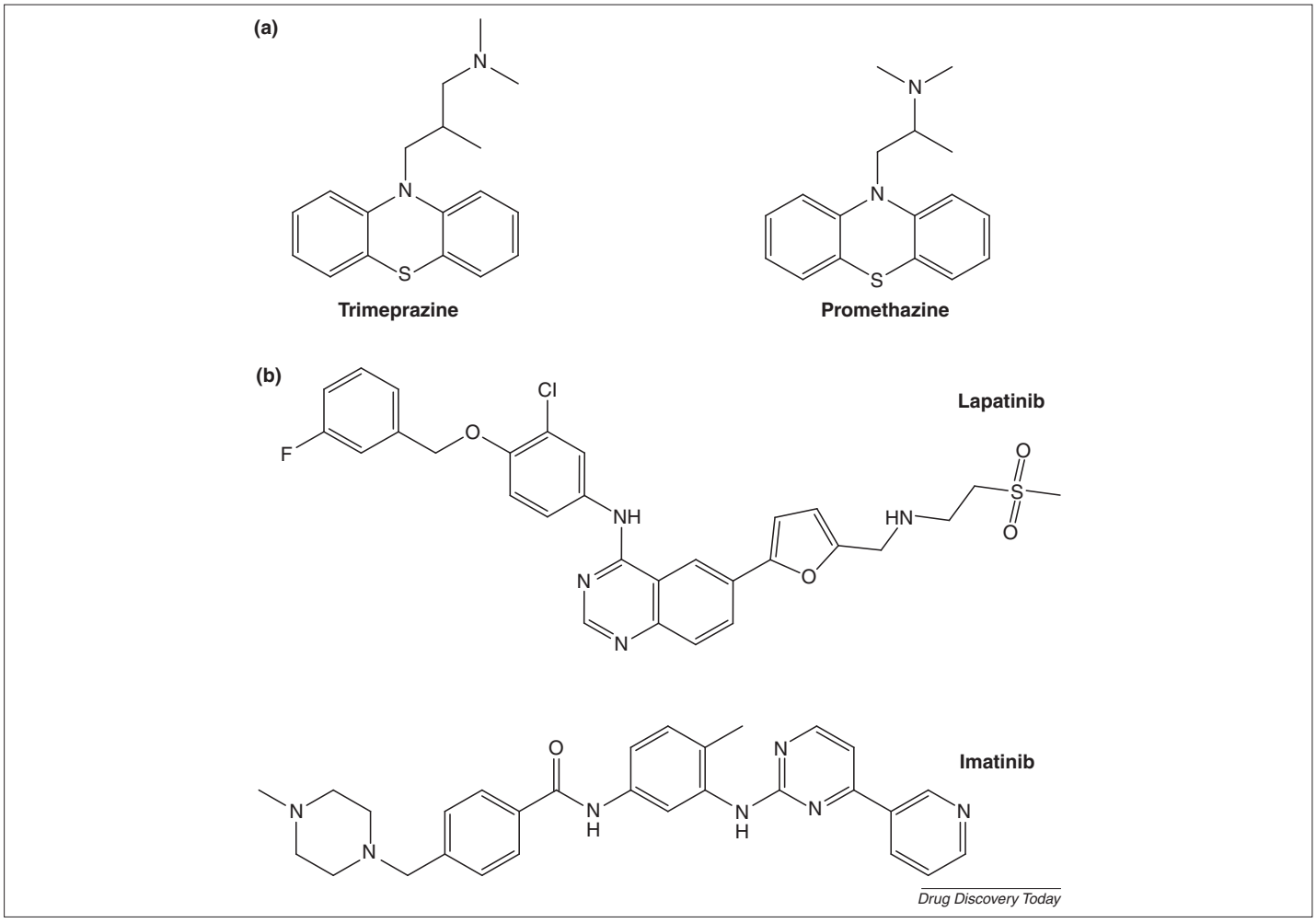


**FIGURE 1**

Approved drugs. Shown are **(a)** two structurally closely related approved drugs, trimeprazine and promethazine, and **(b)** two marketed tyrosine kinase inhibitors, lapatinib and imatinib.

Fig. 1b shows two well-known ATP site-directed protein kinase inhibitors applied in oncology, lapatinib and imatinib. Table 1 reports activity/target annotations for these drugs retrieved from different compound data sources. Despite their high structural similarity, trimeprazine and promethazine are annotated with different numbers of protein targets in DrugBank, two versus 14, respectively. Trimeprazine's primary target is the histamine H1 receptor, whereas promethazine additionally acts on various isoforms of the muscarinic acetylcholine receptor as well as other G protein coupled receptors. Hence, at a first glance, promethazine is a more promiscuous drug than trimeprazine. Here, promiscuity refers to well-defined interactions between an active compound or a drug with multiple targets (rather than nonspecific binding events) [10,11]. So far so good. However, the picture gets much more complicated when we proceed beyond DrugBank and consider activity data from other sources. For example, ChEMBL contains a total of 10 different target annotations for trimeprazine. However, when applying filters for well-defined activity measurements (specifically reported $IC_{50}$ and/or $K_i$ values against human targets) and assays capturing direct ligand–target interactions at highest confidence level, no target annotation remains. BindingDB contains three defined activity measurements for trimeprazine including a $K_i$ value for its primary target. For promethazine, ChEMBL reports a total of 147 different targets. These target annotations are reduced to only 22 after filtering for high-confidence data. From BindingDB, 18 activity annotations are obtained. Hence, DrugBank, ChEMBL and BindingDB report overlapping yet distinct activities and target sets for these drugs. In addition, Open PHACTS contains six pharmacology records for trimeprazine and 113 for promethazine. Furthermore, PubChem reports that trimeprazine was tested in 10 assays and was detected to be active three times, whereas promethazine was tested in 1840 assays with 52 detected activities. Taken together, the data not only corroborate the view that promethazine is a more promiscuous drug than trimeprazine, but also illustrate that it is difficult to differentiate between assay activities and targets. Also, why was promethazine tested in so many more PubChem assays than trimeprazine – and only infrequently found to be active (i.e. in only ~2.8% of all assays), if it is indeed highly promiscuous? Such questions can typically not be answered by analyzing drug data, but they probably would also not be raised without detailed data analysis. Promethazine and trimeprazine are rather typical examples that highlight the heterogeneity and complexity of drug data. Clearly, great care must be taken to analyze target