ELSEVIER

Breaking free from chemical spreadsheets

Matthew Segall, Ed Champness, Chris Leeding, James Chisholm, Peter Hunt, Alex Elliott, Hector Garcia-Martinez, Nick Foster and Samuel Dowling

Optibrium Ltd, 7221 Cambridge Research Park, Beach Drive, Cambridge CB25 9TL, UK

Drug discovery scientists often consider compounds and data in terms of groups, such as chemical series, and relationships, representing similarity or structural transformations, to aid compound optimisation. This is often supported by chemoinformatics algorithms, for example clustering and matched molecular pair analysis. However, chemistry software packages commonly present these data as spreadsheets or form views that make it hard to find relevant patterns or compare related compounds conveniently. Here, we review common data visualisation and analysis methods used to extract information from chemistry data. We introduce a new framework that enables scientists to work flexibly with drug discovery data to reflect their thought processes and interact with the output of algorithms to identify key structure–activity relationships and guide further optimisation intuitively.

Introduction

In drug discovery, project scientists think about their compounds in many different ways. At the level of an individual compound, we want to know its biological and physicochemical properties. For example, is it active against the intended target(s)? Does it have appropriate absorption, distribution, metabolism and excretion (ADME) properties? Is it likely to have off-target effects or cause toxicity? However, although the ultimate goal of every discovery project is the nomination of a high quality development candidate, this outcome is typically the result of investigations of many compounds. Therefore, to help us to navigate this selection and optimisation process, we often organise compounds using a variety of conceptual frameworks. We often consider compounds in groups, such as chemical series, clusters or 'bins' (e.g. progress, reject and study further). We also consider relationships between compounds: optimisation steps or transformations to find modifications that improve activity or other properties; synthetic steps; structure–activity relationships (SAR) that will guide further compound optimisation; and retrospective analysis of project progression in the hope of learning lessons for future projects.

Given the many and varied ways that project scientists consider compounds, their data and relationships, it is perhaps surprising that software packages to support drug discovery chemistry almost always presents those data as spreadsheets or form views (Fig. 1). These are essentially long lists that make it hard to find relevant patterns, focus on subsets of data or even conveniently compare a small number of related compounds. To overcome this constraint, we have even seen project teams print their compounds on sheets of paper and spread them out on a table. Some technological approaches have been explored to address this, for example by Roche (http://youtu.be/3qrQTLs1hPs); but in the age of modern, touch interfaces (Fig. 2), from the perspective of user interaction, chemistry software has largely been stuck in the 1990s.

Often, advanced chemoinformatics algorithms are used to analyse complex compound data and extract important patterns and SAR. However, although these can identify inconspicuous relationships between compounds and their properties, the output also tends to be presented as yet more tables or spreadsheets, making it difficult to interpret and act upon the results and often

Corresponding author: Segall, M. (matt.segall@optibrium.com)

http://dx.doi.org/10.1016/j.drudis.2015.03.008

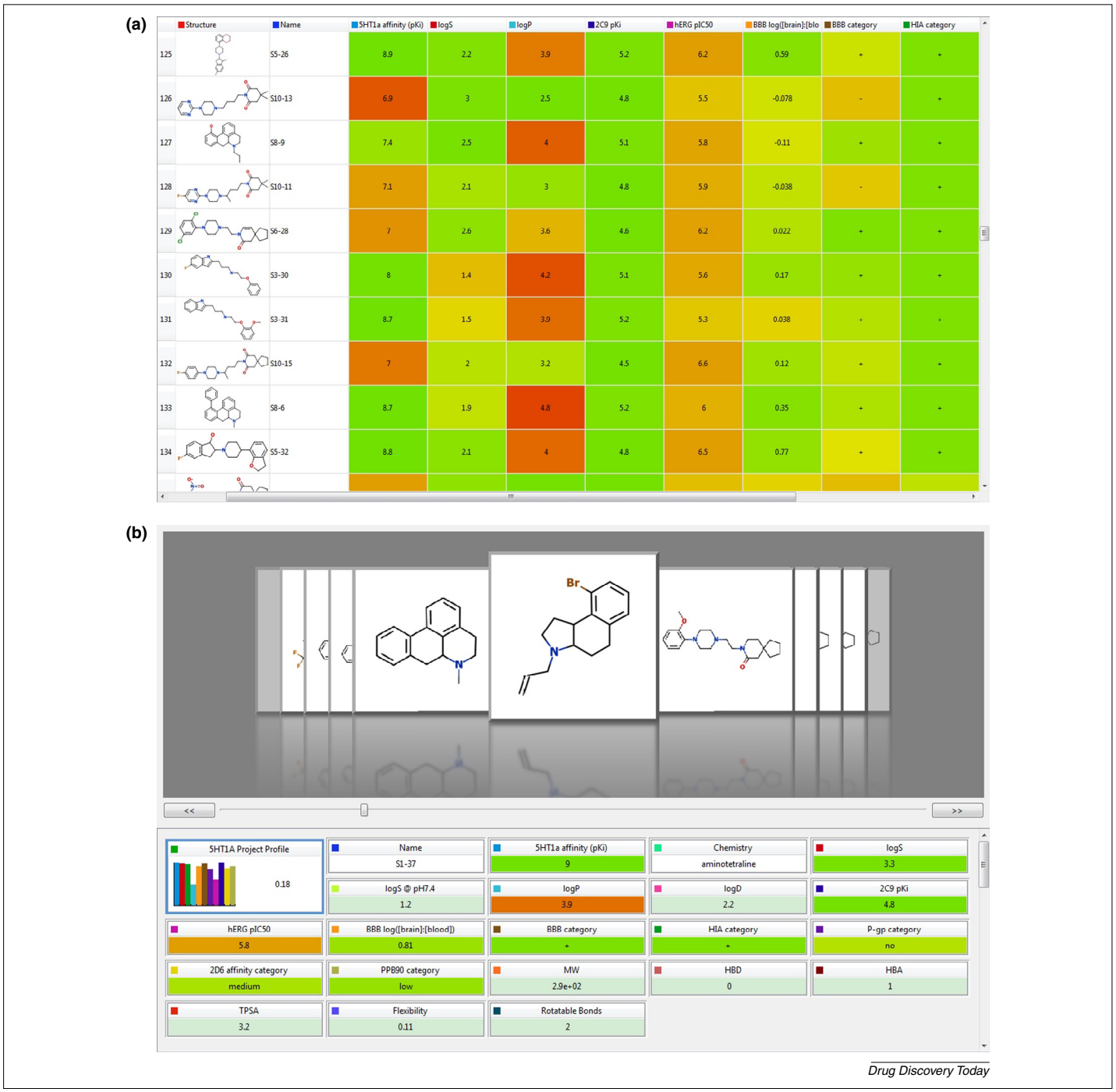

FIGURE 1

Examples of spreadsheet and form views of compound datasets. **(a)** A spreadsheet in which each row represents a compound and the columns contain data including the compound structure and identifier, experimentally measured and calculated properties. The data cells in the spreadsheet have been coloured to produce a heat map on a colour scale from ideal values in green to unacceptably poor values in red. **(b)** An example of a form view in which the properties of a single compound (shown centred) are summarised. The properties for which criteria have been defined are also highlighted in a heat map on the same colour scale as in (a).

needing an expert to analyse the output to reach a conclusion. Data visualisations, such as scatter plots, box plots, pie charts, histograms and SAR tables can help, and illustrative examples are shown in Fig. 3. However, beyond a link to a spreadsheet of data, so that points in a plot can be selected to highlight the corresponding rows, these are static displays of the raw data. They do not allow a scientist to impose their own order on the data to represent the way in which they are thinking about the project compounds. Perhaps paradoxically, we would like to visualise structured data in an unstructured way.

In this review, we will present the methods that are commonly used to impose order on, and extract information from, chemistry data and the ways in which this information is typically viewed. We will then introduce a new framework, in which compound structures and data are arranged as cards, which can be positioned, stacked and linked under the complete control of a scientist. We