New technology to reduce library sizes whilst maintaining diversity

Jo Whelan, freelance writer

A novel computational technology has been developed to optimize the structural diversity of combinatorial screening libraries. The technology, known as ChemSpace, was recently patented in the US by Tripos (St Louis, MO, USA)[1]. Each compound in the library is treated as a collection of molecular fragments (reagents), which can be searched and used to estimate the properties of the compound.

Problems of reducing library sizes

To maximize the likelihood of finding a compound with the required activity in relation to the biological target, the compounds in a combinatorial screening library should, ideally, exhibit as wide a range of chemical and physical characteristics as possible. Recent computational advances have enabled the production of huge virtual libraries in the order of 10^{12} molecules. However, to cut time and costs for both screening and synthesis, there is an urgent need to reduce libraries to a more manageable size without sacrificing their diversity.

Many of the inactive compounds in a conventionally designed library are structurally very similar, giving a high degree of redundancy. Given the principle that structurally similar molecules are expected to have similar biological properties, the new methodology creates an optimally diverse library by selecting one molecule of each structurally similar group present in the universe of possible compounds (Fig. 1). To achieve this, the notion of similarity needs to be quantified using descriptors based on aspects of molecular structure. However, until now, there has been no way to validate these descriptors, i.e. to determine whether similarity in terms of a given descriptor adequately predicts similarity in biological activity. This has severely hampered previous attempts to look for clusters of biologically active compounds within the theoretical universe of compounds. 'Without valid descriptors, the sampling produced from diversity analysis is essentially random, or may even be worse than a random sample – it could be highly skewed,' says Tripos's David Patterson, the chief inventor of the descriptor validation methodology on which the technology is based. 'This technology enables us, for the first time, to design screening libraries and hit follow-up strategies rationally.'

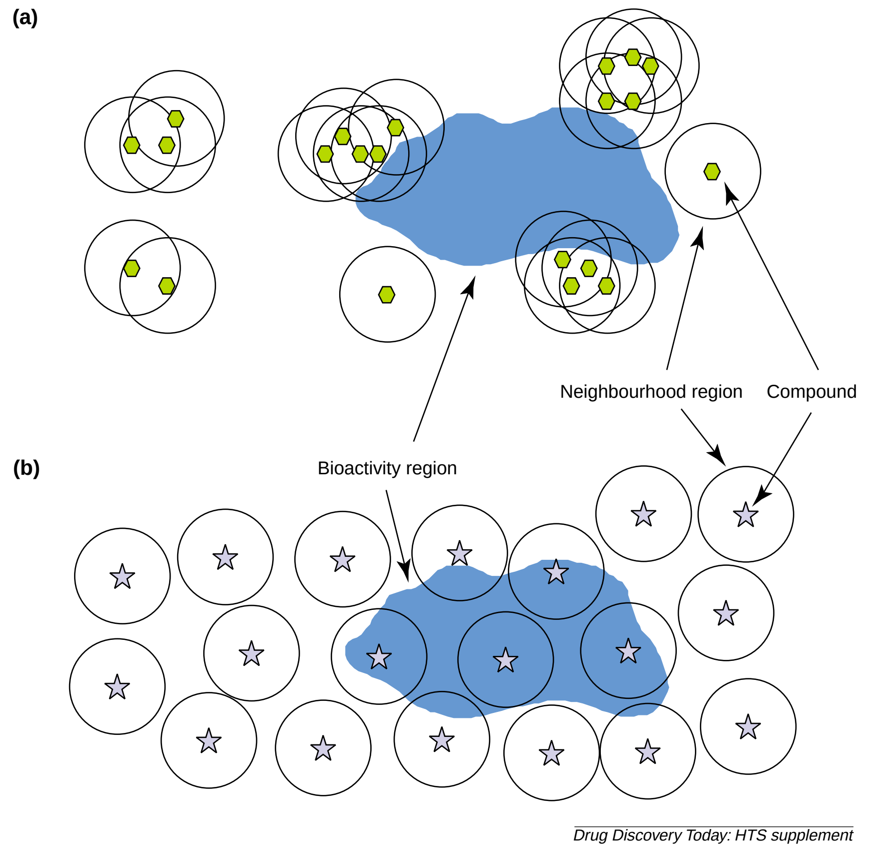

Figure 1. Typical versus designed compound library. (a) In the typical library, many compounds have overlapping neighbourhood regions and are therefore redundant, but a large part of the desirable area (bioactivity region) remains unsampled. (b) In the designed library, neighbourhood regions are spaced out and there is a better chance that the bioactivity region will be included.

Using neighbourhood behaviour

In ChemSpace, validation is based on the concept of 'neighbourhood behaviour'[2] – the ability of a descriptor to group together molecules that have similar biological activity. Patterson and colleagues ranked 11 descriptors by applying them to 20 sets of structure and activity data taken randomly from the literature[2]. Using a computational algorithm, absolute distances between descriptor values for the molecules in the datasets were plotted against the difference in their biological activity. When a descriptor is valid, a small difference in its value does not usually produce a large difference in activity,

1359-6446/01/$ – see front matter ©2001 Elsevier Science Ltd. All rights reserved. PII: S1471-1931(01)00155-0

and this is reflected in the characteristic shape of the resulting graph. The most useful descriptors were found to be a 2D structural fingerprint of the molecule's side chains, and two novel topomeric fields (shape descriptors) known as steric CoMFA (comparative molecular field analysis) fields and H-bonding CoMFA fields[2,3]. Random numbers were used as an example of an invalid descriptor, and gave diffuse plots with no distribution enhancement[2].

Diversity and similarity

In lead discovery it is most efficient to test compounds whose neighbourhood regions do not overlap. However, lead exploration requires the testing of compounds in the same neighbourhood region as the lead. The application of ChemSpace to this problem was addressed in a trial carried out by Tripos and Bristol-Myers Squibb[4]. The structures of four known drugs in the 'sartan' class of angiotensin II (A-II) receptor-antagonists were used as the 'hits'. The aim was to see whether structures identified as their close neighbours were at least as likely to be active as those selected by chemists using traditional concepts of similarity. An A-II-targeted virtual library of 2.6 billion compounds was set up and searched using similarity criteria based on Tripos's topomeric shape descriptor. The 63 closest compounds selected by ChemSpace using similarity criteria were synthesized, as were a further 362 using conventional selection criteria. All were tested for inhibition of A-II receptor binding. Seven compounds were found to be highly active; all were from the group of 63 selected by ChemSpace ($p<0.001$).[4]

According to Tripos, the technology has already successfully achieved results in ongoing projects. For example, a drug discovery collaboration with Arena Pharmaceuticals (San Diego, CA, USA) for novel drugs targeting G-protein-coupled receptors, moved from inception to lead series in under three months and then from lead optimization to *in vivo* biology in less than a further five months.

The future

'The topomer approach seems an excellent way to estimate the similarities of molecules in a biological setting,' says Robert Glen, Professor of Molecular Sciences Informatics at the University of Cambridge, UK. 'In library design and lead follow-up this is undoubtedly an extremely useful idea generator. The most urgent direction for future research is in generating better descriptors, in particular for ADME/Tox.'

Tripos says that the method has been extended to visualize the distribution of compounds in these thousand-dimensional descriptor spaces and to analyze HTS data. It is also being applied in preliminary studies of molecular toxicology and differential protein expression. Continued development of better and more specific descriptors is an ongoing area of research.

References

1 Cramer, R.D. *et al.* [Tripos] (2001) Method for selecting an optimally diverse library of small molecules based on validated molecular structure descriptors. US6185506

2 Patterson, D.E. *et al.* (1996) Neighborhood behaviour: a useful concept for validation of 'molecular diversity' descriptors. *J. Med. Chem.* 39, 3049–3059

3 Cramer, R.D. *et al.* (1996) Biosisosterism as a molecular diversity descriptor: steric fields of single 'topomeric' conformers. *J. Med. Chem.* 39, 3060–3069

4 Cramer, R.D. *et al.* (1999) Prospective identification of biologically active structures by topomer shape similarity searching. *J. Med. Chem.* 42, 3919–3933

Reversing the paradigm for HT protein identification and validation

Rebecca N. Lawrence, Supplements Editor

A new high-throughput technology that promises to rapidly identify and validate new protein drug candidates and targets has been awarded broad patent protection[1]. The technology, developed by Cytos Biotechnology AG (Zurich, Switzerland), is an integrated system for the functional expression of all proteins from a given cell or tissue and subsequent HTS of these proteins for a desired function.

The recent production of the draft sequences of the human genome has led to an explosion in the number of genes discovered. However, the identification of novel functions of genes and proteins has been unable to keep pace due to a lack of appropriate screening technologies. Rapid production of the corresponding purified proteins and effective vectors for expressing these genes in cells or animal models is also necessary for the successful exploitation of the therapeutic potential of the human genome.

Current approaches to identifying gene function start from the gene sequence and, through an elaborate process of *in vitro* and *in vivo* assays, eventually lead to elucidation of the gene function, but this process can sometimes take up to two years. The new technology developed by Cytos works by reversing this paradigm. Claudine Blaser,

1359-6446/01/$ – see front matter ©2001 Elsevier Science Ltd. All rights reserved. PII: S1471-1931(01)00157-4