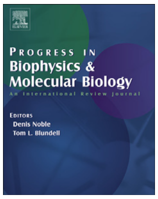
Progress in Biophysics and Molecular Biology 116 (2014) 82–91

Contents lists available at ScienceDirect

Progress in Biophysics and Molecular Biology

journal homepage: www.elsevier.com/locate/pbiomolbio



# Efficient exploration of chemical space by fragment-based screening

Richard J. Hall, Paul N. Mortenson, Christopher W. Murray*

*Astex Pharmaceuticals, 436 Cambridge Science Park, Milton, Road, Cambridge CB4 0QA, United Kingdom*

ARTICLE INFO

*Article history:*
Available online 28 September 2014

*Keywords:*
Fragment-based drug design
Fragment library
Size of chemical space
Properties of fragment hits
3D fragments

ABSTRACT

Screening methods seek to sample a vast chemical space in order to identify starting points for further chemical optimisation. Fragment based drug discovery exploits the superior sampling of chemical space that can be achieved when the molecular weight is restricted. Here we show that commercially available fragment space is still relatively poorly sampled and argue for highly sensitive screening methods to allow the detection of smaller fragments. We analyse the properties of our fragment library versus the properties of X-ray hits derived from the library. We particularly consider properties related to the degree of planarity of the fragments.

## 1. Introduction

Chemical space, taken here to mean the set of organic molecules of suitable size and composition to potentially be oral drugs, is vast. Lipinski et al. have performed an analysis of oral drugs and because 90% were shown to be less than 500 Da, they suggest 500 Da is a useful cut off for the maximal size of a drug-like molecule (Lipinski et al., 1997). This consideration has led to one frequently cited estimate of the size of drug-chemical space that suggests it could consist of well over $10^{60}$ compounds (Bohacek et al., 1996). As others have also observed (Hann and Oprea, 2004), this number is so incomprehensibly large that it bears illustration. Assuming an average molecular mass of between 400 and 500 Da, a sample containing just one molecule of each possible compound would have around $10^{11}$ times the mass of the entire planet. Synthesising a typical 10 mg amount of each would require more matter than is believed to exist in the observable universe.

The subset of chemical space that has been synthesised can be estimated from a snapshot of online chemical databases. As of May 2013, PubChem contains 47 million compounds (PubChem, pubchem.ncbi.nlm.nih.gov), ChemSpider over 28 million compounds (ChemSpider, www.chemspider.com), ZINC contains 21 million compounds (ZINC, zinc.docking.org), ACD contains 7 million compounds (Available Chemicals Directory, http://accelrys.com/products/databases/sourcing/available-chemicals-directory.html), eMolecules 5.9 million (eMolecules, www.emolecules.com) and ChEMBL 1.3 million compounds (Gaulton et al., 2012). It has been estimated that the total number of chemicals that have been synthesised is somewhere in the region of $10^8$ compounds (Renner et al., 2011).

Computational enumeration of the whole of chemical space is a task well beyond current technological capabilities, but if we restrict both the size of molecules and the kinds of chemistry involved, small subsets can be fully explored. This approach leads to more conservative estimates of the size of chemical space, and has been most impressively illustrated by the group of Reymond, with their GDB databases (Blum and Reymond, 2009; Ruddigkeit et al., 2012). The largest of these (GDB-17) contains 166 billion molecules and includes molecules with up to 17 heavy atoms (i.e., non-hydrogen atoms). Analysis of the smaller GDB-13 database suggests that the number of potential molecules increases 8-fold with the addition of each heavy atom, which if extrapolated leads to around $10^{30}$ molecules at 36 heavy atoms (approximately 500 Da). It should be noted however, that the chemical rules employed in these enumerations are fairly restrictive, and fail to generate many molecules present in public databases.

An *in silico* exploration of $10^{30}$ compounds also looks like an insurmountable challenge for the foreseeable future; storing such a compound set would require of the order of $10^{23}$ gigabytes using the relatively compact InChI format (Heller et al., 2013). By comparison, based on a recent study, a reasonable estimate of the world's data storage capacity might be of the order of $10^{12}$ or $10^{13}$ gigabytes (Hilbert and López, 2011). Possibly a larger problem would be generating the structures – if we allow ourselves a generous $10^8$ s (approximately 3 years) for this exercise, we will still need to generate $10^{22}$ structures per second.

Rather than trying to exhaustively enumerate chemical space, some researchers have tried to describe it in various ways (Bemis and Murcko, 1996; Oprea and Gottfries, 2001; Pollock et al., 2008;

* Corresponding author. Tel.: +44 1223 226228.
*E-mail address:* chris.murray@astx.com (C.W. Murray).

http://dx.doi.org/10.1016/j.pbiomolbio.2014.09.007
0079-6107/© 2014 Elsevier Ltd. All rights reserved.

Reymond and Awale, 2012; Schuffenhauer et al., 2006; Schwartz et al., 2013; Virshup et al., 2013). These descriptive approaches can help to visualise, partition and categorise chemical space but will not be considered any further in this article.

## 2. Fragments as probes of chemical space

A commonly stated advantage of fragment-based screening is that a given number of fragment-sized molecules can sample chemical space much more efficiently than the same number of larger molecules. But what do we actually mean when we talk about sampling chemical space? Most obviously, we can consider what fraction of the available chemical space we have represented in a screening collection/library. If we limit the screening library to compounds of MW $< 500$, then as discussed above the total chemical space is at least $10^{30}$ compounds, meaning that even the largest multimillion compound screening library represents only an infinitesimal fraction of what is possible. However, if we are much more restrictive and limit ourselves to smaller molecules, we can start to populate more meaningful fractions of the virtual set. For example, in GDB-13 there are around 100 million compounds with a heavy atom count of 12, meaning that a selection of just 1000 compounds represents 0.001% of the whole of that subset of chemical space. This is 19 orders of magnitude better than the fractional coverage achieved by our much larger library of drug-sized compounds.

An alternative but perhaps more relevant interpretation (from a drug discovery perspective) would be to recast the concept of covering chemical space as a question: How many molecules do I need in my library to ensure a sufficient number of hits against an arbitrary target? This question was analysed theoretically in 2001, in a seminal paper by Hann and co-workers (Hann et al., 2001). In this article, they constructed a simple model of protein-ligand binding, whereby both protein and ligand consist simply of one-dimensional strings of binary features. These features might represent shape, hydrophobic, or electrostatic properties of the underlying molecules, or indeed any aspect of them that needs to match in order for binding to occur. In the model, any mismatch between a ligand feature and protein feature will prevent binding, so a successful binding event requires matching all of the ligand features with complementary features in the protein. Hann and co-workers constructed libraries of model ligands with varying numbers of these features, and virtually screened them against sets of model receptors. They found that the hit rates of these libraries decreased as the number of ligand features increased; more complex molecules are less likely to bind to any given target. Hann and co-workers also recognised that smaller ligands have fewer features so, in general, will bind less tightly and the observed hit rate will depend on the sensitivity of the method used to detect the hits. Due to these two competing effects, the probability of detecting a binding event for a random ligand of a specific size is predicted to be: (i) low for very small ligands (due to the limits of sensitivity); (ii) highest for small ligands (reflecting the balance between the two competing effects); (iii) monotonically decreasing for larger ligands (due to the increasing probability of a mismatch).

More recently, Leach and Hann revisited this analysis (Leach and Hann, 2011), and discussed several attempts to validate it using real experimental data. The results were in general equivocal, but there were a number of complications in interpreting the data. Firstly, in the absence of a universally recognised and well-behaved descriptor of molecular complexity, many studies focussed on size, either in the form of molecular weight or heavy atom count. While in the original model system, size correlates perfectly with complexity, for real molecules this will not be the case. Highly functionalised molecules with ornate three dimensional shapes are clearly more complex than equivalently sized molecules that are planar and less decorated. Moreover, there is the confounding factor of lipophilicity, which tends to correlate with size, at least in collections of molecules associated with drug discovery. Lipophilicity tends to correlate positively with promiscuity, which might mask the expected opposite correlation with complexity. In addition, most of the studies examined looked at promiscuity of compounds, defined using a certain potency threshold; the choice of this threshold (which varied between studies) will obviously have a potentially large impact on the conclusions drawn.

Another group performed an experimental analysis that did match the predictions of the Hann model (Teotico et al., 2009). They performed a virtual fragment screen for inhibitors of β-lactamase using a docking methodology, and obtained 23 hits from 48 compounds tested. This compared very favourably with hit rates from previous HTS and virtual screening campaigns. Many of the 23 fragment hits represented chemotypes not seen before as inhibitors of this enzyme, so the researchers looked to see if these chemotypes were present in commercial databases of lead-like compounds (defined here as molecules with a heavy atom count (HAC) $\leq 25$). Although they found 675 compounds matching the fragment chemotypes, most of these were analogues of only 2 of the fragment hits. In addition, using GDB-11 (Fink and Reymond, 2007) to provide side chain groups, they estimated the number of possible lead-like molecules containing the 23 fragment hits to be around $10^{11}$. By contrast, considering fragment-sized molecules (HAC $\leq 17$), they found 93 commercial compounds out of a theoretical $10^4$. The authors hypothesised that this improved coverage of the available chemical space at lower heavy atom counts was a large factor in their improved hit rate.

## 3. Coverage of fragment space by commercial vendors

Any analysis of fragment space is frustrated by (a) the inability to enumerate chemical space for all but the smallest fragments; and (b) the inability to prevent the generation of enormous numbers of unsuitable compounds. Here we address the first problem by considering a well-defined subspace of chemical space that can be easily enumerated, and we discuss the second problem at the end of this section. Our chemical subspace is composed of simple fragment topologies based on mono-substituted and di-substituted six membered rings. Rings of this type represent the more common rings observed in drug molecules. The rings are substituted by one or two side chains with each side chain containing up to 6 atoms. A graph theoretical method was used to exhaustively generate all possible side chains as described below. Databases of commercial compounds were then compared to the enumerated library, to ascertain which of these topologies was present.

Here we describe the construction of the side chain library that was subsequently attached to the six membered ring. All connected graphs of two to seven nodes were exhaustively generated using the program GENG (McKay, 1981). These were filtered to remove any graph containing more than one cycle. The resulting graphs were represented as smiles strings by converting each node to an $sp^3$ carbon and each edge to a single bond. From the resulting set of 70 graphs, a side chain library was generated by systematically replacing each singly connected atom with an attachment point that could subsequently be referenced in a virtual "SMIRKS reaction" that attaches the side chain to a core scaffold. Any graph that did not contain a singly connected atom (e.g. the graph that would match cycloheptane) was discarded. After removal of duplicates this procedure led to the creation of 96 graphs, each of which contained the attachment point and up to 6 atoms that would form part of the product of the SMIRKS reaction. These are shown in Fig. 1.