Library design for NMR-based screening

Christopher A. Lepre

Recent advances in NMR-based screening methods have made it possible to screen larger libraries of molecules with higher throughput. However, experience shows that intelligent library design is important if NMR screening is to succeed in aiding our discovery of potent and useful lead compounds. This review presents the current state-of-the-art methodologies for designing primary and follow-up libraries for NMR screening. Diversity, drug-likeness and combinatorial libraries are discussed, and the inherent pitfalls of the NMR approach are addressed.

Christopher A. Lepre
Vertex Pharmaceuticals
130 Waverly Street
Cambridge, MA 02139-4242
USA
tel: +1 617 577 6627
fax: +1 617 577 6645
e-mail: lepre@vpharm.com

▼ NMR spectroscopy has long been used to detect the binding of small molecules to biomolecular targets, but has only recently been employed to screen libraries of compounds for drug discovery. A variety of well known NMR techniques have been used for screening, including nuclear Overhauser effect (NOE)[1–4], chemical shift perturbation[5], diffusion[6–8], relaxation[9] and saturation transfer[10–12]. NMR screening can be used to detect binders with affinities ranging from nanomolar to millimolar, and recent method improvements have drastically reduced protein consumption and increased throughput.

Despite the burgeoning interest in NMR screening, there has been little discussion in the literature of how best to build libraries of compounds for this purpose. Thoughtful library design is essential, not only to avoid complications in the NMR experiments, but also to avoid wasting time and reagents, and to provide information that is both meaningful and directs the search towards more potent and desirable lead compounds.

Based upon our own experience and anecdotal evidence from other industrial laboratories, NMR screens fail for a variety of reasons:

- Primary screens can fail to produce hits if the library is not sufficiently large or diverse, if the compounds are insoluble or aggregated, or if the detection threshold is too insensitive (e.g. if the ligand:protein ratio is too high).
- A high fraction of false positives can result if the compounds are insoluble or aggregated, or if non-specific binding occurs. The latter seems to occur most often when the protein construct being screened is a domain from a larger protein and contains surface-exposed hydrophobic patches or partially unfolded regions.
- Follow-up libraries based on NMR screening results can fail to produce potent hits if they are too small or unimaginative (e.g. based on simple substructure searching without allowing for sufficient structural variation). Alternatively, over-interpretation of structure–activity relationships (SAR) for primary screening hits can result in an overly focussed, misdirected follow-up library.
- When NMR screening produces hits, the synthesis of analogs cannot be pursued unless the compounds are synthetically accessible. This is particularly true when NMR and conventional HTS are run against the same target because the NMR hits are typically less potent.

Many of these problems can be avoided by careful design of the NMR screening libraries. This review describes the application of four principal criteria for library design: diversity, drug-like character, solubility and synthetic accessibility, and is not intended to provide an in-depth coverage of general library design principles, which are instead reviewed in Refs 13–18.

Diversity

In designing screening libraries, if the system to be screened exhibits ‘neighborhood behavior’

1359-6446/01/$ – see front matter ©2001 Elsevier Science Ltd. All rights reserved. PII: S1359-6446(00)01616-0

(i.e. similar molecules have similar biological activities), it can be assumed that a diverse collection of compounds is inherently superior to a random set. Neighborhoods are defined as the regions surrounding each compound in the multidimensional space defined by various molecular descriptors[19]. Testing the activity of the compound in the center of a neighborhood is thought to predict that of other molecules within the region. The most efficient approach is therefore to avoid overlaps of neighborhood regions by spreading out the molecules in descriptor space (i.e. maximize diversity).

Diversity is not an inherent property; it depends on the specific descriptors used. The natural preference is to use descriptors that have been shown to correlate with biological activity (e.g. fingerprints or whole-molecule descriptors)[13,20]. The computational methods for measuring molecular diversity have been extensively reviewed[13,15,21,22], and commercially available software allows the non-specialist to use methods such as similarity ranking, and hierarchical and non-hierarchical clustering[23,24].

Although diversity is desirable, it is not particularly useful as the sole criterion for library design. Indeed, in some cases little is gained by maximizing diversity. When compound space is large or the neighborhoods are small, each data point covers little space. Overlap is unlikely because relatively few compounds are screened and, regardless of what selection method is used, approximately the same amount of space is covered. In other cases, neighborhood behavior might not be obeyed; the medicinal chemistry literature contains numerous examples of small modifications to molecules that dramatically affect activity. In addition, it has been shown that pure diversity biases libraries away from desirable drug properties. Combinatorial libraries designed chiefly for diversity tend to produce hits that are large, highly flexible, insoluble and lipophilic, whereas orally available drugs are typically smaller and only moderately flexible or lipophilic[25].

In simple terms, the need for diversity is inversely proportional to the knowledge that exists for that target [Ref. 27; Baldwin, R (1996) *Conference on Libraries and Drug Discovery*, 29–31 January, Coronado, CA, USA]. For lead discovery, the strategy is to search as widely as possible within the set of compounds that possess the preferred characteristics for leads. In this process we focus, filter and diversify: first, focus upon desirable classes of molecules for inclusion; second, filter out compounds with unsuitable physicochemical characteristics; and third, maximize diversity. For lead follow-up, the library is focussed on the neighborhood of the initial hits, and information regarding the structure of the target or an active pharmacophore is used to prioritize the compounds.

Drug-like character

A popular approach for designing screening libraries is to bias the selection of compounds towards drug-like molecules. This approach is based on the assumption that compounds resembling known drugs are more likely to possess desirable biological properties such as low toxicity, high oral absorption and permeability, resistance to metabolic degradation, and the absence of rapid excretion. Despite recent advances in predicting solubility, absorption and permeability[27–29], most methods published to date have relied on retrospective analysis of known drugs to recognize the drug-like compounds in a pool of candidate molecules. A disadvantage of this approach is the reduced likelihood of discovering radically different classes of drugs. In particular, if the non-drug-like comparison set contains as yet undiscovered drugs, this approach might direct the search away from these novel classes.

Several groups have compared chemical and drug databases to identify molecular descriptors that can be used to classify molecules as drug-like or non-drug-like[25,30,31]. Some have also tried to incorporate medicinal chemistry experience by including descriptors based on molecular fragments or local structures that: (1) occur in top-selling drugs; (2) were found in previous assay hits; or (3) were preferred by chemists[25,32]. Others have used neural networks to classify compounds as drug-like or non-drug-like[33–36]. In general, methods based on whole-molecule descriptors should be used cautiously if the goal is to include small fragments in the library that are representative of larger drug molecules. Classification schemes that have been defined using intact drug molecules might fail to recognize highly novel molecules[32] or small fragments as being drug-like.

A simple rule for classifying compounds on the basis of their likelihood to be orally available was devised at Pfizer (Groton, CT, USA)[37]. This frequently cited ‘Rule-of-5’ (Box 1) limits the molecular weight (MW), logP and number of hydrogen bond donors and acceptors to values commonly found in orally available compounds. Other physicochemical properties, such as the number of rings, heavy atoms and rotatable bonds, have also been used to predict drug-likeness (Box 1)[38,39]. The use of rotatable bonds as a parameter is considered important because of the entropic penalty that accompanies the binding of flexible molecules. A set of 25 rules for drug-like compounds was recently published by researchers at Boehringer Ingelheim (Ridgefield, CT, USA)[40]. Finally, the percentage of saturation of daylight fingerprints has been used to identify and remove both simple (<10% saturated) and complex (>60% saturated) molecules[41].

In contrast to other methods that classify compounds, an appealing approach is to construct a library around

Box 1. Properties used to select drug-like compounds

‘Rule of 5’ criteria[a]

- Molecular weight ≤500 Da
- LogP ≤5
- Hydrogen bond donors (OH and NH) ≤5
- Hydrogen bond acceptors (O and N) ≤10

Other criteria[b–d]

- Number of heavy atoms 10–70
- Rotatable bonds 2–8
- Number of rings 1–6, aromatic ≤3
- Molar refractivity 40–130

References

[a] Lipinski, C.A. *et al*. (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* 23, 3–25

[b] Ghose, A.K. *et al.* (1999) A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. *J. Comb. Chem.* 1, 55–68

[c] Oprea, T.I. (2000) Property distribution of drug-related chemical databases. *J. Comput. Aided Mol. Des.* 14, 251–264

[d] Xu, J. and Stevenson, J. (2000) Drug-like index: a new approach to measure drug-like compounds and their diversity. *J. Chem. Inf. Comput. Sci.* 40, 1177–1187

molecular scaffolds that occur frequently in known drugs[4,42]. Shape description methods have been used to show that, surprisingly, only 32 simple molecular frameworks describe approximately half of the drugs in the Comprehensive Medicinal Chemistry (CMC) database[43]. When atom type, hybridization and bond order are included, a quarter of the drugs contain one of only 41 frameworks. The most common drug side chains were similarly identified[44]. These results suggest that a few common, drug-like scaffolds and side chains can be used to design a general-purpose drug-like screening library.

Many companies routinely use filters to remove compounds containing chemical moieties that are insoluble, chemically reactive or correlated with toxicity. Researchers at Amgen (Thousand Oaks, CA)[45], GlaxoWellcome (Stevenage, UK)[41] and Rhône-Poulenc Rorer (Dagenham, UK)[46] have published examples of such empirical functional group filters. Alternatively, a web-accessible program for ‘Rapid Elimination Of Swill’ (REOS) that contains filters based on physical properties and on over 200 functional groups can be used[47]. However, such filtering methods must be applied intelligently or they can lead to rejection of useful compounds. For example, we find that when the REOS filter is blindly applied to known drugs from the CMC database, it rejects 73% of them. Others have reported similar problems with overly conservative drug-like filters wrongly classifying known drugs[14,32].

Despite the emphasis on drug-like molecules, the goal of primary screening is to produce good leads, not drugs. As described by Teague and coworkers[48], drug-like leads pass Lipinski’s ‘Rule of 5’, but have molecular weights (350–500 Da) and clogP values (3–5) at the high end of the preferred range. Such leads are not the best starting points for optimization because addition of lipophilic groups to increase potency can adversely affect their pharmacokinetic (PK) properties. Lead-like leads are smaller (MW=100–350 Da) and more polar (clogP 1–3), so addition of lipophilic groups often improves both their potency and PK properties[48]. In addition, these small molecules can still bind avidly, particularly when they possess a single charge. This is not surprising because even molecules with as few as 10–20 atoms are known to bind to proteins with nanomolar affinities[49]. The implication is that libraries biased towards small, polar molecules are inherently superior. This is advantageous for NMR screening, which benefits from both high solubility and relatively weak binding.

Solubility

Because of the inherently low sensitivity of NMR experiments, NMR screening requires compounds with much higher aqueous solubility than do conventional screening methods. Reported concentrations for compounds used in screening experiments range from 1–10 mM for NOE and chemical shift-based or affinity-NMR screening methods[1,4–8,50], to ~50 μM for more sensitive one-dimensional methods[9–12] or experiments using a cryoprobe[51].

It would be very useful to be able to prescreen for solubility at the library design stage. Although many empirical approaches have been published, there is currently a lack of facile and accurate methods to predict solubility for diverse sets of compounds. One well known method[52] uses MW, melting temperature (T_m), octanol–water partition coefficient (K_{ow}) and correction factors for certain structural features[53]. However, for most compounds, experimental values for T_m and K_{ow} are unavailable, and the accuracy of the prediction is reduced when values for these parameters are estimated.

As a result, at present there is no good alternative to experimentally confirming the solubility of all compounds before screening. This is essential because even optically clear solutions can contain high-MW aggregates of compounds. Such aggregates can produce false negatives as a result of co-aggregation with protein and removal of the compounds from solution, or false positives owing to high rotational correlation time (τ_c) of the aggregate, which is misinterpreted as a hit in many ligand-based detection schemes. Even protein-based detection methods (i.e. chemical shift perturbation) are subject to interference as a result of aggregation[54].

Mixture design

If compounds are to be screened in mixtures, then deconvolution of hits, competition and interactions between the components must also be considered. Mixtures must be deconvoluted to identify the hits when using protein-based detection methods. The total number of experiments required to screen and deconvolute the mixtures is at a minimum when the number of compounds per mixture is approximately equal to $1/(\text{hit rate})^{1/2}$ (A. Ross, pers. commun.). For a 10% hit rate, the optimal number of compounds in a mixture is only three, so it might be preferable simply to screen individual compounds[54].

Competition between mixture components becomes a potential problem when a significant fraction of the compounds bind to the target. The probability of getting *x* hits from *n* compounds can be derived from a binomial distribution (Eqn 1), where *p* is the probability of getting a hit from one compound:

$$f(x) = \left(\frac{n!}{x!\,(n-x)!}\right)p^{x}(1-p)^{n-x} \qquad [1]$$

Thus, given a 10% hit rate, 26% of mixtures of ten compounds will have two, three or four competing hits, whereas only 2.7% of mixtures of three will contain two hits.

Compounds in mixtures can also react with one another. This is a serious problem because the identity of the species binding to the target becomes questionable. Hann and coworkers[41] report that for random mixtures of ten compounds stored in dimethyl sulfoxide (DMSO), approximately 25% showed evidence of chemical reactions between the components after three months. Carefulmixture design to separate acids from bases and electrophiles from nucleophiles reduced the incidence of reactions to 9%.

The likelihood of reaction between components depends on the number of possible pair-wise interactions (*npairs*), which increases dramatically with mixture size *n* (Eqn 2):

$$npairs = \left(\frac{n-1}{2}\right)n \qquad [2]$$

Thus, a mixture of three compounds has three possible pair-wise interactions, a mixture of ten compounds has 45, and a mixture of 100 compounds has 4950. Although some researchers have reported screening 100 compounds in a mixture[51], it is expected that intercomponent reactions would occur in most mixtures of 100, many of them involving multiple compounds. Therefore, the most prudent approach is to confirm the integrity of all mixtures before screening.

Synthetic accessibility

The preferences of the end users (e.g. chemists and modelers) concerning the information that will be obtained from NMR screening must be considered. Most compounds that chemists regard as particularly undesirable are removed in the filtering steps, but drug-likeness alone is not enough to warrant a compound's inclusion in the library. Because NMR hits are often relatively weak (high micromolar to millimolar), they will elicit little interest unless they are also synthetically accessible.

Chemists typically prioritize screening hits on the basis of 'drug-like feel' and synthetic accessibility, intellectual property potential and potency. Analogs are then made of molecules that appear to fall into distinct classes. The goal is the early and efficient identification of lead classes that exhibit promising SAR. In this context, molecules that require multiple steps or low-yield reactions to make analogs are unattractive.

Opinions regarding which molecules are synthetically attractive can differ from one chemist to the next. To address this issue a library was constructed from compounds containing linkers and side chains that are accessible using reactions implemented by our combinatorial chemistry group. Linkers are defined as atoms on the direct path connecting two ring systems, and side chains are non-ring, non-linker atoms (i.e. functional groups)[43]. The compounds in the 'SHAPES linking library' contain either the scaffolds used in the original SHAPES library[42] or proprietary scaffolds, along with either a preferred linker or side chain. All of the compounds pass through the REOS filter, and are thus ostensibly drug-like in character. Any hits produced by this library are highly amenable to follow-up using conventional chemistry or combinatorial methods.

Design of follow-up libraries

A particular strength of NMR screening is its ability to drive a 'decision tree' approach, in which information about simple binding scaffolds directs the search towards more elaborate and potent compounds in a systematic and convergent manner. Small, simple molecules are desirable because they can penetrate deep into active sites without steric hindrance. In addition, the number of compounds necessary to represent the scaffolds of interest is relatively small. As additional scaffolds are linked together, the number of compounds that must be screened to represent all possible combinations increases geometrically – a familiar problem in combinatorial chemistry.

Once hits have been found, the first goal in designing a follow-up library is to search the chemical neighborhood and identify SARs. This is usually done by introducing substituents that explore chemical space and can make additional interactions with the target, thereby increasing potency. The second goal is to make non-trivial modifications to the scaffold without disrupting the 'bindacophore'. Ideally, one would like to escape a local minimum in chemical space to find a more potent scaffold class in a nearby neighborhood, or 'scaffold hop'[55] (i.e. identify isofunctional molecules with significantly different structures). Finally, bioisosteric replacements[56,57] can be sought for functional groups that correlate with activity.

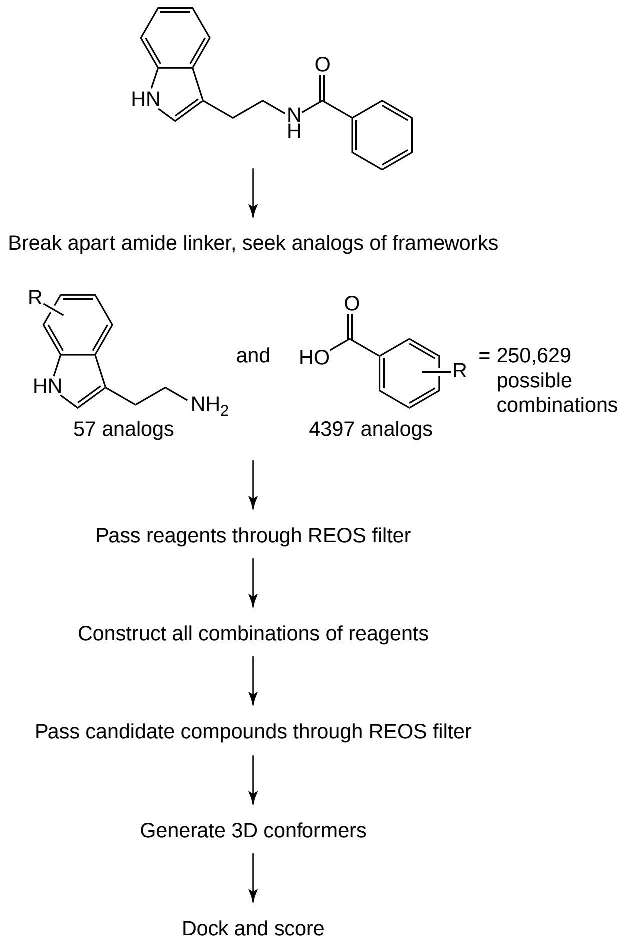
Searches for commercially available compounds in the neighborhood of a hit are readily carried out using fragment-based and similarity searches[58] (routinely available in commercial database software) or nearest-neighbor searches such as nearest-neighbor cluster analysis[59]. To allow for greater structural variation and to identify less intuitively obvious analogs, searches can also be carried out using pharmacophore-based descriptors[17,60,61]. The challenge for this approach is correctly aligning the hits to define the pharmacophore, which in the absence of atomic-level information about the ligand–protein complex, might be impossible. Fortunately, small and soluble NMR screening compounds are readily absorbed into protein crystals, and structures can sometimes be rapidly obtained even for compounds with millimolar binding affinities.

There are several general strategies by which information about weakly binding molecular fragments can be used to construct more potent molecules. One is a fragment-linking approach, in which an attempt is made to link together molecules that individually bind to the target. This can be done by 'fragment fusion', which is the blind testing of compounds that contain two or more scaffolds known to bind[4]. Alternatively, in 'SAR by NMR'[5], fragments that bind in proximity are identified using chemical shift perturbation and NOE experiments, and linked compounds are then synthesized. 'Combinatorial target-guided ligand assembly' screens fragments that contain a common linking group, and all possible combinations of binding fragments are then combinatorially synthesized[62]. A drawback of these methods is that, without detailed information about the bound conformations of the individual fragments, it is possible that the linked analogs will not be correctly oriented to bind to the protein.

An alternative strategy is to systematically elaborate upon the initial hits using an iterative process. Information from each round of screening is used to bias the selection of compounds for the next round of increasingly complex compounds. We have observed for several targets that HTS screens of follow-up libraries based on NMR screening hits produce hit rates tenfold higher than do general HTS screens, thereby demonstrating convergence towards more potent compounds. An interesting application of a convergent strategy to the design of DNA gyrase inhibitors starting from millimolar hits was recently reported[63]. In this case, the lead optimization process employed SAR information and structure-based drug design.

Another strategy is to combinatorially permute scaffolds shown to bind by NMR. In addition to varying the functional groups attached to a combinatorial scaffold (for which examples abound in the literature), hits can be retrosynthetically fragmented into building-block fragments that are then systematically permuted[64]. If the structure of the target active site is known, or can be reliably modeled based on homologous proteins, then computational methods can be used to guide the design of follow-up compounds to be synthesized. These 'virtual screening' methods have been applied to select compounds for combinatorial synthesis from vast hypothetical libaries[65–70]. NMR screening results should increase the likelihood of success by restricting the virtual screen to scaffolds that have been experimentally proven to bind to the target of interest. In addition, by narrowing the number of putative scaffolds, it is possible to screen all combinations of scaffolds rather than being forced to arbitrarily reduce the number of virtual compounds to a computationally manageable number.

A virtual screening strategy for designing follow-up libraries based on linked hits is shown in Fig. 1. To design analogs of screening hits containing frameworks joined by a linker, the bound molecule is conceptually broken apart at the linker. A list of possible molecular fragments to replace each framework is then created using available reagents, and this list is edited using the REOS filter. Next, the molecules for all possible combinations of these fragments are derived and REOS-filtered. This second filtering step is necessary because although individual fragments pass the REOS filter, the combining of fragments creates some candidate compounds that fail (e.g. those which exceed the maximum charge or hydrogen bond donor–acceptor cutoffs). Although REOS filtering could be left until the end of the process, it is applied earlier to reduce the computational expense of the combination step. Three-dimensional conformers are then generated, docked into the active site using a genetic-algorithm-based method to find the best binding site, and scored[47]. One advantage of starting with hits that contain scaffolds that are already linked (rather than individual fragments), is that the scaffolds are already correctly oriented for binding.


HN
O
N
H
Break apart amide linker, seek analogs of frameworks
R
HN
NH2
57 analogs
and
O
HO
R
4397 analogs
= 250,629 possible combinations
Pass reagents through REOS filter
Construct all combinations of reagents
Pass candidate compounds through REOS filter
Generate 3D conformers
Dock and score

Drug DiscoveryToday

Figure 1. Virtual screening strategy for following up SHAPES linking library hits. A hypothetical NMR screening hit consisting of two linked scaffolds (top) is followed up by identifying commercially available analogs for these scaffolds. Putative combinations of these reagents are passed through successive REOS and virtual screening filters. Abbreviation: REOS, rapid elimination of swill.

Building the SHAPES linking library

The construction of the SHAPES linking library illustrates how the principles described in this review can be applied. In the 'focus' step of the process, the decision was made to start with a set of 1.2 million commercially available compounds and focus upon those containing scaffolds and/or side chains commonly found in known drugs. Of the 41 most common drug frameworks[43], 13 were not used because of their synthetic complexity, inherently low solubility or lack of commercially available analogs. These frameworks included steroids, tetracyclines, opiates and β-lactam antibiotics. Additional frameworks were added to represent scaffolds that were not among the 41 most common drug frameworks, but did occur frequently in known drugs and were of particular interest (e.g. were synthetically attractive).

In the 'filter' step, approximately 60% of the candidates were eliminated by the REOS filter (Fig. 2). Then, compounds that contained combichem-accessible linkers or side chains (17%) were selected using a substructure search of SMILES (Simplified Molecular Input Line Entry System)[71] strings. Approximately 25% of those compounds were predicted to be soluble either using the Syracuse Research predictor[52] or on the basis of clogP values ≤3.

The object of the 'diversify' step was to optimize the diversity while reducing the number of compounds to a number that could be readily screened. Jarvis-Patrick clustering[24] was used to cluster the candidates, and 8% of the compounds from the filtering step were manually selected from the centroids of the clusters, thereby making use of the fact that with the Jarvis-Patrick method, the centroids of the clusters are representative of the library as a whole. An advantage of selecting centroids instead of singletons is that additional compounds can be ordered from the same cluster to follow up any hits. Several rounds of cascade clustering[72] were applied to reduce the number of singletons and small clusters. Compounds that were overly elaborate or highly symmetric, or that contained multiple hydroxyl groups (which tend to bind proteins non-specifically, e.g. glycerol), were avoided.

Of the compounds that were ordered from commercial vendors, ~50% actually arrived in time to be included in this version of the library, and ~50% of these were determined to be soluble to 1 mM in aqueous buffer using NMR spectra, thereby leaving ~500 compounds in the final library.

Conclusions

In summary, the focus, filter and diversify procedure is generally useful for constructing libraries of compounds for screening by enzymological or physically based methods. The selection criteria presented here tailor such libraries to the particular needs of NMR screening, and enhance the value of the resulting hits. Complementary combinatorial approaches allow NMR screening hits to be rapidly exploited by synthesis of targeted libraries.

Ongoing advances in instrumentation and experimental methods are changing the role of NMR screening in drug discovery programs. Increases in sensitivity have dramatically reduced the amount of protein required, allowing NMR screening to be carried out earlier in the lifetime of a drug discovery project. Thus, NMR might prove to be a universal assay that is of particular value when run before development of an enzymatic screen. For example, once small quantities of a new target have been obtained, an NMR screen can be run to determine its 'drugability' and

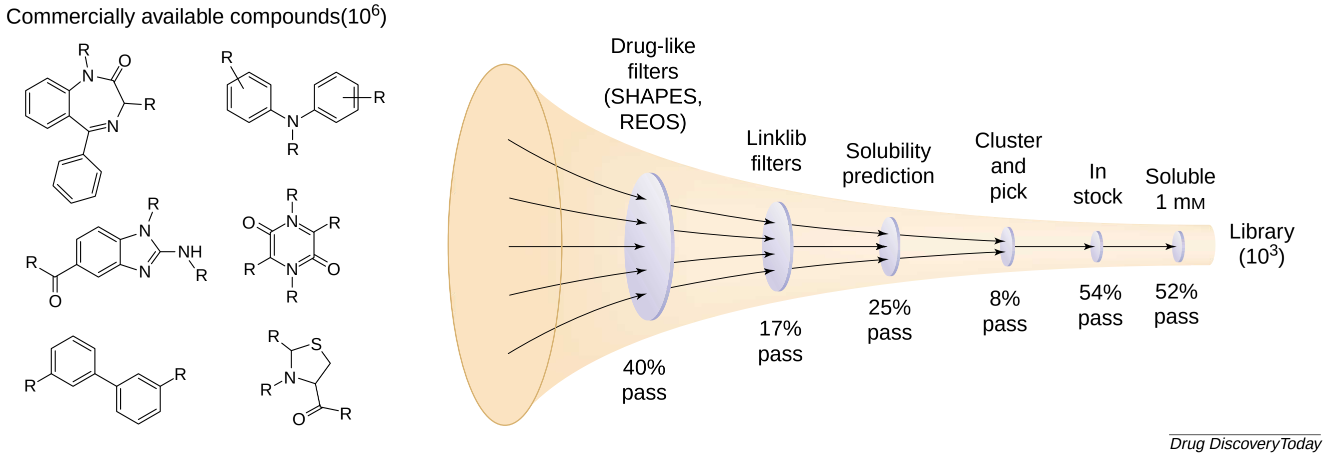


Figure 2. Building the SHAPES linking library. Commercially available compounds are successively filtered on the basis of drug-like character, the presence of synthetically accessible linkages and functional groups, predicted solubility, clustering, availability and actual solubility. Percentages indicate the fraction of compounds that successfully pass through the individual stages in the process.

hence decide whether it is worth the expense of scaling up protein production and developing an enzymatic assay for that target. In addition, early knowledge about which scaffolds bind to a new target helps to establish an intellectual property position, as well as directing preparation of follow-up libraries so that they are ready by the time an HTP enzymological screen is developed. Lastly, NMR screening can be used to help develop an HTP assay by identifying potential probe molecules.

The use of genomic information is stimulating interest in screening libraries that are focussed upon families of related proteins. Libraries of proprietary scaffolds are especially valuable because of the competitive advantage they offer. Because of its ease of use and low protein consumption, NMR screening might be employed to screen leads for one target against libraries of related proteins, thus broadening the patent coverage of a proprietary lead class.

As advances in NMR methods allow the screening of increasing numbers of compounds, the intelligent design of diverse, drug-like, NMR-compatible and synthetically accessible libraries will continue to be of crucial importance.

Acknowledgements

The author wishes to thank Jonathan Moore, Mark Murcko and Pat Walters for helpful discussions, as well as Brian Goldman, Marie Lepre and Stephan Ogenstad for assistance with the statistical analysis.

References

1 Meyer, B. *et al.* (1997) Screening mixtures for biological activity by NMR. *Eur. J. Biochem.* 246, 705–709

2 Chen, A. and Shapiro, M.J. (1998) NOE pumping: a novel NMR technique for identification of compounds with binding affinity to macromolecules. *J. Am. Chem. Soc.* 120, 10258–10259

3 Chen, A. and Shapiro, M.J. (2000) NOE pumping. 2. A high-throughput method to determine compounds with binding affinity to macromolecules by NMR. *J. Am. Chem. Soc.* 122, 414–415

4 Fejzo, J. *et al.* (1999) The SHAPES strategy: an NMR-based approach for lead generation in drug discovery. *Chem. Biol.* 6, 755–769

5 Shuker, S.B. *et al.* (1996) Discovering high-affinity ligands for proteins: SAR by NMR. *Science* 274, 1531–1534

6 Lin, M. and Shapiro, M.J. (1996) Mixture analysis in combinatorial chemistry. Application of diffusion-resolved NMR spectroscopy. *J. Org. Chem.* 61, 7617–7619

7 Lin, M.L. *et al.* (1997) Diffusion-edited NMR – affinity NMR for direct observation of molecular interactions. *J. Am. Chem. Soc.* 119, 5249–5250

8 Lin, M. *et al.* (1997) Screening mixtures by affinity NMR. *J. Org. Chem.* 62, 8930–8931

9 Hajduk, P.J. *et al.* (1997) One-dimensional relaxation- and diffusion-edited NMR methods for screening compounds that bind to macromolecules. *J. Am. Chem. Soc.* 119, 12257–12261

10 Mayer, M. and Meyer, B. (1999) Characterization of ligand binding by saturation transfer difference NMR spectroscopy. *Angew. Chem., Int. Ed. Engl.* 38, 1784–1788

11 Klein, J. *et al.* (1999) Detecting binding affinity to immobilized receptor proteins in compound libraries by HR-MAS-STD NMR. *J. Am. Chem. Soc.* 121, 5336–5337

12 Dalvit, C. *et al.* (2000) Identification of compounds with binding affinity to proteins via magnetization transfer from bulk water. *J. Biomol. NMR* 18, 65–68

13 Drewry, D.H. and Young, S.S. (1999) Approaches to the design of combinatorial libraries. *Chemometr. Intell. Lab. Syst.* 48, 1–20

14 Leach, A.R. and Hann, M.M. (2000) The *in silico* world of virtual libraries. *Drug Discov. Today* 5, 326–336

15 Van Drie, J.H. and Lajiness, M.S. (1998) Approaches to virtual library design. *Drug Discov. Today* 3, 274–283

16 Higgs, R.E. *et al.* (1997) Experimental designs for selecting molecules from large chemical databases. *J. Chem. Inf. Comput. Sci.* 37, 861–870

17 Good, A.C. and Lewis, R.A. (1997) New methodology for profiling combinatorial libraries and screening sets: cleaning up the design process with HARPick. *J. Med. Chem.* 40, 3926–3936

18 Cramer, R.D. *et al.* (1998) Virtual compound libraries: a new approach to decision making in molecular discovery research. *J. Chem. Inf. Comput. Sci.* 38, 1010–1023

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

20 Bayada, D.M. *et al.* (1999) Molecular diversity and representivity in chemical databases. *J. Chem. Inf. Comput. Sci.* 39, 1–10

21 Gorse, D. *et al.* (1999) Molecular diversity and its analysis. *Drug Discov. Today* 4, 257–264

22 Mason, J.S. and Hermsmeier, M.A. (1999) Diversity assessment. *Curr. Opin. Chem. Biol.* 3, 342–349

23 Ward, J.H. (1963) Hierarchical grouping to optimize and objective function. *J. Am. Stat. Assoc.* 58, 236–245

24 Jarvis, R.A. and Patrick, E.A. (1973) Clustering using a similarity measure based on shared near neighbors. *IEEE Trans. Comput.* C-22, 1025–1034

25 Martin, E.J. and Critchlow, R.E. (1999) Beyond mere diversity: tailoring combinatorial libraries for drug discovery. *J. Comb. Chem.* 1, 32–45

26 Hann, M. and Green, R. (1999) Chemoinformatics – a new name for an old problem? *Curr. Opin. Chem. Biol.* 3, 379–383

27 Blake, J.F. (2000) Chemoinformatics – predicting the physicochemical properties of 'drug-like' molecules. *Curr. Opin. Biotechnol.* 11, 104–107

28 Clark, D.E. and Pickett, S.D. (2000) Computational methods for the prediction of 'drug-likeness'. *Drug Discov. Today* 5, 49–58

29 Stenberg, P. *et al.* (2000) Virtual screening of intestinal drug permeability. *J. Control. Release* 65, 231–243

30 Cummins, D.J. *et al.* (1996) Molecular diversity in chemical databases: comparison of medicinal chemistry knowledge bases and databases of commercially available compounds. *J. Chem. Inf. Comput. Sci.* 36, 750–763

31 Gillet, V.J. *et al.* (1998) Indentification of biological activity profiles using substructure analysis and genetic algorithms. *J. Chem. Inf. Comput. Sci.* 38, 165–179

32 Wang, J. and Ramnarayan, K. (1999) Toward designing drug-like libraries: a novel computational approach for prediction of drug feasibility of compounds. *J. Comb. Chem.* 1, 524–533

33 Ajay, *et al.* (1998) Can we learn to distinguish between 'drug-like' and 'nondrug-like' molecules? *J. Med. Chem.* 41, 3314–3324

34 Ajay, *et al.* (1999) Designing libraries with CNS activity. *J. Med. Chem.* 42, 4942–4951

35 Sadowski, J. and Kubinyi, H. (1998) A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* 41, 3325–3329

36 Sadowski, J. (2000) Optimization of chemical libraries by neural networks. *Curr. Opin. Chem. Biol.* 4, 280–282

37 Lipinski, C.A. *et al.* (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 23, 3–25

38 Ghose, A.K. *et al.* (1999) A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. *J. Comb. Chem.* 1, 55–68

39 Oprea, T.I. (2000) Property distribution of drug-related chemical databases. *J. Comput. -Aided Mol. Design* 14, 251–264

40 Xu, J. and Stevenson, J. (2000) Drug-like index: a new approach to measure drug-like compounds and their diversity. *J. Chem. Inf. Comput. Sci.* 40, 1177–1187

41 Hann, M. *et al.* (1999) Strategic pooling of compounds for high-throughput screening. *J. Chem. Inf. Comput. Sci.* 39, 897–902

42 Peng, J.W. *et al.* NMR-based approaches for lead generation in drug discovery. *Methods in Enzymol.* (in press)

43 Bemis, G.W. and Murcko, M.A. (1996) The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* 39, 2887–2893

44 Bemis, G.W. and Murcko, M.A. (1999) Properties of known drugs. 2. Side chains. *J. Med. Chem.* 42, 5095–5099

45 Rishton, G.M. (1997) Reactive compounds and *in vitro* false positives in HTS. *Drug Discov. Today* 2, 382–384

46 Lewis, R.A. *et al.* (1997) Similarity measures for rational set selection and analysis of combinatorial libraries: the diverse property-derived (DPD) approach. *J. Chem. Inf. Comput. Sci.* 37, 599–614

47 Walters, W.P. *et al.* (1998) Virtual screening – an overview. *Drug Discov. Today* 3, 160–178

48 Teague, S.J. *et al.* (1999) The design of lead-like combinatorial libraries. *Angew. Chem., Int. Ed. Engl.* 38, 3743–3747

49 Kuntz, I.D. *et al.* (1999) The maximal affinity of ligands. *Proc. Natl. Acad. Sci. U. S. A.* 96, 9997–10002

50 Gonnella, N. *et al.* (1998) Isotope-filtered affinity NMR. *J. Magn. Reson.* 131, 336–338

51 Hajduk, P.J. *et al.* (1999) High-throughput nuclear magnetic resonance-based screening. *J. Med. Chem.* 42, 2525–2517

52 Meylan, W.M. *et al.* (1996) Improved method for estimating water solubility from octanol–water partition coefficient. *Environ. Toxicol. Chem.* 15, 100–106

53 Meylan, W.M. and Howard, P.H. (1995) Atom/fragment contribution method for estimating octanol–water partition coefficients. *J. Pharm. Sci.* 84, 83–92

54 Ross, A. *et al.* (2000) Automation of NMR measurements and data evaluation for systematically screening interactions of small molecules with target proteins. *J. Biomol. NMR* 16, 139–146

55 Schneider, G. *et al.* (1999) Scaffold-hopping: by topological pharmacophore search: a contribution to virtual screening. *Angew. Chem., Int. Ed. Engl.* 38, 2894–2896

56 Kubinyi, H. (1998) Similarity and dissimilarity: a medicinal chemist's view. *Perspect. Drug Discov. Des.* 9/10/11 , 225–252

57 Lipinski, C.A. (1986) *Bioisosterism in drug design* (Vol. 21), (Allen, R.C., ed.), Academic Press

58 Willet, P. *et al.* (1998) Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* 38, 983–996

59 Stanton, D.T. *et al.* (1999) Application of nearest-neighbor and cluster analysis in pharmaceutical lead discovery. *J. Chem. Inf. Comput. Sci.* 39, 21–27

60 Pickett, S.D. *et al.* (2000) Enhancing the hit-to-lead properties of lead optimization libaries. *J. Chem. Inf. Comput. Sci.* 40, 263–272

61 Marriott, D.P. *et al.* (1999) Lead generation using pharmacophore mapping and three-dimensional database searching: application to muscarinic M3 receptor antagonists. *J. Med. Chem.* 42, 3210–3216

62 Maly, D.J. *et al.* (2000) Combinatorial target-guided ligand assembly: identification of potent subtype-selective c-Src inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* 97, 2419–2424

63 Boehm, H-J. *et al.* (2000) Novel inhibitors of DNA gyrase: 3D structure based needle screening, hit validation by biophysical methods, and 3D guided optimization. *J. Med. Chem.* 43, 2664–2674

64 Lewell, X.Q. *et al.* (1998) RECAP – Retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* 38, 511–522

65 Böhm, H-J. *et al.* (1999) Combinatorial docking and combinatorial chemistry: design of potent non-peptide thrombin inhibitors. *J. Comput. -Aided Mol. Design* 13, 51–56

66 Böhm, H-J. and Stahl, M. (2000) Structure-based library design: molecular modelling merges with combinatorial chemistry. *Curr. Opin. Chem. Biol.* 4, 283–286

67 Perola, E. *et al.* (2000) Successful virtual screening of a chemical database for farnesyltransferase inhibitor leads. *J. Med. Chem.* 43, 401–408

68 Julián-Ortiz, J.V. *et al.* (1999) Virtual combinatorial syntheses and computational screening of new potential anti-herpes compounds. *J. Med. Chem.* 42, 3308–3314

69 Schapira, M. *et al.* (2000) Rational discovery of novel nuclear hormone receptor antagonists. *Proc. Natl. Acad. Sci. U. S. A.* 97, 1008–1013

70 Li, J. *et al.* (1998) Targeted molecular diversity in drug discovery: integration of structure-based design and combinatorial chemistry. *Drug Discov. Today* 3, 105–112

71 Weininger, D. (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* 28, 31–36

72 Menard, P.R. *et al.* (1998) Rational screening set design and compound selection: cascaded clustering. *J. Chem. Inf. Comput. Sci.* 38, 497–505