J Comput Aided Mol Des (2011) 25:611–620
DOI 10.1007/s10822-011-9461-x

Design of a Fragment Library that maximally represents available chemical space

M. N. Schulz · J. Landström · K. Bright · R. E. Hubbard

Received: 4 June 2011 / Accepted: 13 July 2011 / Published online: 27 July 2011

Abstract Cheminformatics protocols have been developed and assessed that identify a small set of fragments which can represent the compounds in a chemical library for use in fragment-based ligand discovery. Six different methods have been implemented and tested on Input Libraries of compounds from three suppliers. The resulting Fragment Sets have been characterised on the basis of computed physico-chemical properties and their similarity to the Input Libraries. A method that iteratively identifies fragments with the maximum number of similar compounds in the Input Library (Nearest Neighbours) produces the most diverse library. This approach could increase the success of experimental ligand discovery projects, by providing fragments that can be progressed rapidly to larger compounds through access to available similar compounds (known as SAR by Catalog).

Keywords Fragment-based ligand discovery · Library design · SAR by catalog · Nearest neighbours

Introduction

There has been considerable activity and development in the methods of fragment-based discovery over the past 15 years. The pioneering work of the Abbott group (SAR by NMR [1, 2]) has been adopted and adapted by many groups to provide an effective method for identifying initial compounds that interact with a target (for recent reviews see [3–5]). The central premise in fragment based drug discovery is to identify the starting point for hit generation by screening a relatively small number (500–2,000) of compounds of low MW (100 Da < MW < 250 Da). Although such small compounds bind quite weakly (typically with dissociation constant, K_D, between 100 and 10 mM), they can be highly ligand efficient [6]. There are three main components to a fragment-screening platform: (a) a method for detecting which fragments bind; (b) methods for developing the fragments to hit compounds and (c) a fragment library.

Detection of binding

Fragments are small, weakly binding hits, with useful fragments with ligand efficiency of 0.3 [6] being found as 150 Da compounds (12 heavy atoms, defined as non-hydrogen atoms) binding with affinities (K_D) as weak as 5 mM. It is challenging for most biochemical screening assays to detect such weak binding, particularly as the high concentration of ligand can disrupt the assay. For this reason, the success of fragment based discovery has driven the development of (and heavily relied upon) a range of biophysical methods for detecting fragment binding, such as ligand-observed and protein-observed NMR [7], X-ray crystallography [8], surface plasmon

Electronic supplementary material The online version of this article (doi:10.1007/s10822-011-9461-x) contains supplementary material, which is available to authorized users.

M. N. Schulz · J. Landström · K. Bright · R. E. Hubbard (✉)
YSBL, Chemistry Department, University of York,
Heslington, York YO10 5DD, UK
e-mail: rod@ysbl.york.ac.uk

R. E. Hubbard
HYMS, University of York, Heslington, York YO10 5DD, UK

R. E. Hubbard
Vernalis (R&D) Ltd, Granta Park, Abington,
Cambridge CB21 6GB, UK

resonance (SPR) [9], capillary electrophoresis followed by mass spectrometry and thermal shift analysis (TSA—also known as differential scanning fluorimetry [10]). The different methods have different constraints on the nature and physico-chemical properties of the fragments and target to detect binding. There has not been a comprehensive comparison of these different techniques against the same target with the same fragment library. It can be difficult to configure a method for a particular target (for example, it may not crystallise with an accessible binding site or insufficient protein cannot be generated for ligand-observed NMR experiments. The different methods can give different numbers and types of false positive and false negative hits. However, experience suggests that if the limitations and constraints of the different methods are understood, then the experimental techniques can identify the same hits [11]. The main constraint on the fragment library for all the experimental procedures is that it is soluble, stable and non-reactive at the high concentrations required (typically 1 mM aqueous) to detect such weak binding.

Methods of progression

There are three main approaches to evolution of fragments to hits with the affinity necessary to register in biochemical assays, such that they can then be progressed using conventional medicinal chemistry techniques [12]. The initial SAR by NMR approach emphasised linking fragments together [2]. However, it can be challenging to identify suitable chemistry to link two fragments together [13] while preserving key interactions from the fragments. More generally used has been combining information from various fragments and virtual screening hits together (so-called fragment merging [14–16]) or fragment growing [17, 18].

A particularly powerful (and generally accessible) approach to the initial growth of fragments is the so-called SAR by catalog approach. The features of a fragment hit are used in a substructure or similarity search of a database of accessible compounds (commercially available, corporate collection or proposed synthetic libraries) to find Nearest Neighbour (NN) compounds as seen in examples from PDPK1 [15] and Hsp90 [19]. Although care must be taken that the NNs are not carrying excessive unwanted functional groups, this can be a rapid and economic way to establish SAR against a target without the need for extensive synthetic chemistry resources. A similar idea is found in the Virtual Fragment Linking approach [20] in which similarity to fragment hits is used to prioritise selection of compounds for assay from a corporate compound collection.

Library design

Fragment libraries have been designed by many different groups from academia as well as large [21] and small [22] pharmaceutical companies. Some recent reviews summarise what has been published [23, 24], however the details of many libraries remain proprietary. The general approach is to identify compounds of suitable molecular weight and physico-chemical properties (number of hydrogen bond donors and acceptors, rotatable bonds, ring systems, calculated solubility) and to exclude any compounds that have functionalities linked to reactivity or toxicology, such as described in one of the earlier papers. The libraries can be generic [25] or target biased, based on privileged fragments from known drug compounds [26], fragments from compounds that have activity against therapeutically relevant targets [27, 28] or a pharmacophore search or virtual screening against the target [29]. In addition, some have designed libraries with particular features required for the detection mechanism such as ^{19}F containing compounds for ^{19}F-NMR based screening [30] or bromine containing compounds for X-ray crystallographic detection [31].

The aim of this paper

In this paper we describe our approach to identify a set of fragments that maximally represents the available compounds and therefore is a Fragment Set that is suitable for the SAR by catalog approach. Input Libraries of compounds from suppliers are filtered to remove unwanted functionality, duplicates and salts and then split into Fragment and Non-Fragment Libraries. A number of different cheminformatics approaches are then used to identify Fragment Sets that contain compounds from the Fragment Library that maximally represent the chemical space of the compounds in the Non-Fragment Library. The characteristics of the compounds in the Fragment Sets are then compared.

Methods

All protocols were implemented in Pipeline Pilot (Student Edition 6.1.5.0 and Professional version 8.0.1.500, http://www.accelrys.com).

Preparation of Input Libraries

Compounds were downloaded from the ZINC database [32] from three suppliers: 600,220 compounds from Asinex (http://www.asinex.com), 79,228 from Maybridge (http://www.maybridge.com) and 321,371 from Specs (http://www.specs.net). The Input Libraries were prepared as

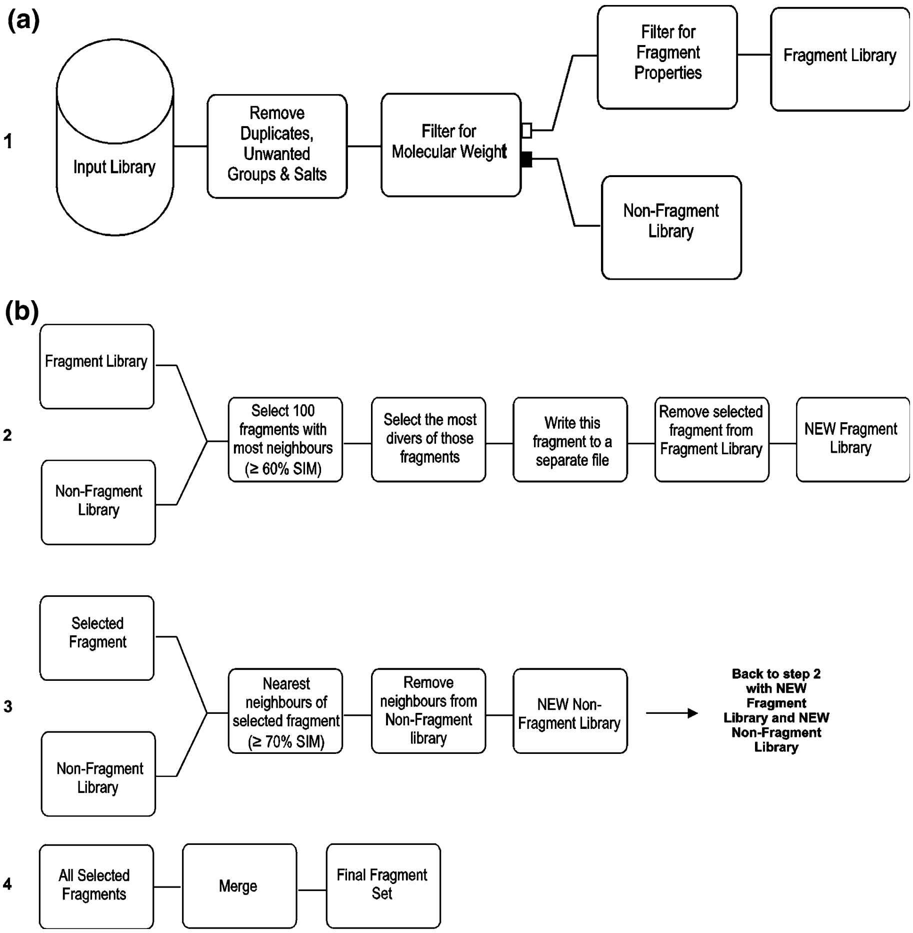
(a)
1
Input Library
Remove Duplicates, Unwanted Groups & Salts
Filter for Molecular Weight
Filter for Fragment Properties
Fragment Library
Non-Fragment Library
(b)
2
Fragment Library
Non-Fragment Library
Select 100 fragments with most neighbours (≥ 60% SIM)
Select the most divers of those fragments
Write this fragment to a separate file
Remove selected fragment from Fragment Library
NEW Fragment Library
3
Selected Fragment
Non-Fragment Library
Nearest neighbours of selected fragment (≥ 70% SIM)
Remove neighbours from Non-Fragment library
NEW Non-Fragment Library
Back to step 2 with NEW Fragment Library and NEW Non-Fragment Library
4
All Selected Fragments
Merge
Final Fragment Set

Fig. 1 Schematic of the iterative removal protocol. (*1*) Preparation of Fragment and Non-Fragment Libraries from Input Library, (*2*) Identification of most diverse fragment with most neighbours, (*3*) Removal of neighbours from Non-Fragment Library, (*4*) Completion

summarised in step 1 of Fig. 1. Duplicates were removed based on canonical SMILES, with every molecule following the first occurrence discarded and all salts removed) were removed from the libraries. In addition, the SMARTS strings (http://www.daylight.com)) listed in Table S1 in Supplementary Information were used to identify compounds for removal that contain unwanted chemical functionality. This generated an Input Library for each supplier. The following steps were applied to process each Input Library. The number of compounds resulting from each step is summarised in Table 1.

Separation into Fragment and Non-Fragment Libraries

For each Input Library, the compounds with MW > 250 Da (>300 Da if Sulphur containing) were assigned to a Non-Fragment Library (total number of compounds). The Fragment Library contains all compounds with MW ≤ 250 Da (≤300 Da if Sulphur) with compounds removed which did not satisfy the following criteria: MW ≥ 100 Da; ≤6 O and N atoms; ≤3 hydrogen bond donors; ALogP ≤ 3.0; ≤3 rotatable bonds, a calculated molecular solubility (determined as LogS) ≥−3.4 and a molecular polar surface area of ≤80 Å^2.

Selecting a Fragment Set from a Fragment Library

The selection is based on different protocols which rely on methods for assessing similarity. In all the calculations reported here, the FCFP_4 (Functional class extended-connectivity fingerprint of maximum diameter 4) as implemented in Pipeline Pilot was used to characterise each compound.

Similarity was assessed using a form of the Tversky equation [33] as:

$$\mathrm{SIM_{Tversky}} = \mathrm{SA}/(\mathrm{SA} + \beta * \mathrm{SB} + \gamma * \mathrm{SC})$$

where SIM is the similarity coefficient, SA is the number of bits defined in both the Fragment and Non-Fragment compound, SB is the number of bits defined in the Fragment but not in the Non-Fragment and SC is the number of bits defined in the Non-Fragment but not in the Fragment. β and γ are weighing factors. Setting $\beta = 1$ and $\gamma = 1$ results in the widely used Tanimoto coefficient

Table 1 The number (and percentage) of compounds remaining in preparation of Fragment and Non-Fragment Libraries from the Input Libraries

Input Library	Asinex	Maybridge	Specs
Initial from ZINC database	600,220	79,228	321,371
Duplicates removed	5,330 (1%)	2,322 (3%)	3,965 (1%)
Salts removed	0	0	0
Unwanted functionality removed	317,755 (53%)	38,812 (49%)	196,331 (62%)
With MW < 250 Da (<300 Da with S)	24,914 (4%)	9,512 (12%)	17,620 (5%)
Removed with properties filter	13,450 (54%)*	5,050 (53%)*	8,829 (50%)*
Removed with solubility filter	3,768 (15%)*	1,434 (4%)*	3,150 (17%)*
Fragment Library	7,696 (1%)	6,484 (8%)	5,641 (2%)
Non-Fragments Library	252,221 (42%)	28,582 (36%)	103,455 (32%)

* Percentage calculated relative to number passing the MW filter

($\mathrm{SIM}_{\mathrm{Tanimoto}}$) which is suitable for comparison of molecules of the same size:

$$\mathrm{SIM}_{\mathrm{Tanimoto}} = \mathrm{SA}/(\mathrm{SA} + \mathrm{SB} + \mathrm{SC})$$

In the analyses presented in this paper, we set $\beta = 1$ and $\gamma = 0$. This leads to a value of 1 when a fragment is wholly present in a larger compound which is a more appropriate value to consider when gauging how much of a fragment is present in a larger compound:

$$\mathrm{SIM}_{\mathrm{SuperstructureTversky}} = \mathrm{SA}/(\mathrm{SA} + \mathrm{SB})$$

$\mathrm{SIM}_{\mathrm{SuperstructureTversky}}$ will be referred to as the Tversky similarity in this paper.

The clustering was performed with a partitioning maximum dissimilarity method using the Tanimoto coefficient and the FCFP_4 fingerprints, i.e. the set of compounds is divided into ever smaller subsets based on a maximum dissimilarity method. In this approach, a random molecule is chosen and named the cluster centre. The most dissimilar to this molecule forms the next cluster centre. The next cluster centre is formed by the molecules the most distant to both chosen centres and so forth until sufficient cluster centres are selected. The other molecules are assembled around these cluster centres. The selection of the most diverse compound is also based on the same maximum dissimilarity method.

Six protocols for each of the three selected suppliers were developed, generating six different Fragment Sets. Table S2 contains diagrams representing the different protocols and links for download. The description below is for use of these protocols to generate Fragment Sets each of 200 compounds from each Input Library.

The first protocol is called "Cluster All". For each compound in a Fragment Library, the average is calculated of the Tversky similarity to all the compounds in the corresponding Non-Fragment Library. The two Libraries (Non-Fragment and Fragment) are then combined and clustered to give 200 clusters. For each cluster, the compound from the Fragment Library with the highest average similarity is selected for the Fragment Set. There can be clusters that do not contain a fragment, so this protocol can give less than the desired 200 compounds in the Fragment Set.

The second protocol is called "Cluster Fragments" and is a variant of Cluster All. The same calculation is performed to calculate the average Tversky similarity to Non-Fragments for each compound in the Fragment Library. In this case, only the Fragment Library is clustered into 200 clusters and the compound with the highest average similarity is taken for the Fragment Set.

The third variant is called "Similarity within Cluster" (short: "SIMwithinCluster"). Here, the Fragment and Non-Fragment Libraries are combined and clustered into 200 clusters. The fragment with the most similarity to the ClusterCenter is selected. For technical reasons (the Pipeline Pilot implementation), this particular selection is based on the Tanimoto similarity.

The approaches called Substructure used two different mechanisms for calculating similarity (Count or Map). The Fragments are broken down into substructures (Rings, rings assemblies, bridge assemblies (ring systems linked by two or more bonds), chains and Murcko assemblies) and the number of times the substructures within a fragment occur in the compounds in the Non-Fragment Library is counted. Two different protocols were used. The "Substructure Count" protocol counts the total number of occurrences of an indicated substructure whereas the "Substructure Map" protocol counts only whether a substructure is present. The two approaches give different results so are treated separately.

The final method is called "Iterative Removal" and was the most complex to implement. Figure 1 (steps 2–4) shows the logic of this protocol. In step 2, the number of Non-Fragments which have a Tversky similarity >60% to each Fragment is counted. The 100 Fragments with the highest number of such "Near-Neighbours" are then compared and the most diverse compound selected for the Fragment Set and removed from the Fragment Library. In step 3, the compounds with a Tversky similarity >70% to this selected Fragment are removed from the Non-Fragment Library. The calculation is then repeated (back to step 2) with the reduced Fragment and Non-Fragment Libraries to identify the next most diverse Fragment with a high number of Near-Neighbours and so on to eventually generate the Fragment Set of 200 compounds (step 4). The similarity cut-offs were chosen to maximise the number of Non-Fragments represented by each member of the Fragment Set, while retaining sufficient Non-Fragments for subsequent selections. Preliminary calculations showed that 50% Tversky similarity removes too many Non-Fragments and 80% too few.

Profiling the Fragment Sets

The aim of the selection protocols is to generate a Fragment Set which represents the maximum number of Nearest Neighbours in the Non-Fragment Library. The Fragment Sets were profiled for physico-chemical properties using the methods implemented in Pipeline Pilot (see Supplementary Information for details). In addition, the Fragment Sets were profiled as:

1. The number of Nearest Neighbours with greater than 70% Tversky similarity (SIMILARITY)
2. The chemical diversity of the Fragment Set, calculated as (1) DIVERSITY 1: the average Tanimoto similarity within the Fragment Set and (2) DIVERSITY 2: how equally distributed the Fragment Set is when clustered on fingerprint, calculated as the standard deviation of the number of compounds in each cluster. In the results reported here, 20 clusters were used as there were 200 fragments.
3. The average similarity for each Fragment with drug-like molecules, calculated as (1) the average similarity with compounds from the Non-Fragments Library that satisfy Lipinski rule of 5 (DRUG-LIKE 1) and (2) the average similarity with compounds from the World Drug Bank (www.drugbank.ca) (DRUG-LIKE 2).
4. The average number of Non-Fragments that have <70% Tversky similarity for each fragment (NON-SIMILAR)

Results and discussion

Table 1 summarises the number of compounds from each supplier (Specs, Maybridge and Asinex) that remained after each stage of processing. (Table S1 also shows the number of compounds removed in generating the Input Library from each supplier for each of the unwanted chemical functionalities).

A large number of compounds are removed by the unwanted functionality filters. The filters for nitro, methylene, tertiary or quarternary amines, acrylates and atoms that are not C, N, F, CL or S remove the most compounds. There is a similar profile of exclusions across the three suppliers, except the Maybridge Input Library has proportionately more acrylate and tertiary amine containing compounds. In addition, the Maybridge Input Library contains a proportionately larger number of compounds (12%) with low MW, compared to Specs (5%) and Asinex (4%), but about the same percentage (50%) of these low MW compounds from each supplier do not have the desired fragment properties. Table S3 lists the number of low MW compounds that do not pass the different properties filters to be selected for the Fragment Library from the low MW list. The main failures are for polarity (ALogP and PSA), with similar percentages of compounds failing for each of the suppliers.

The properties of the resulting Fragment Libraries are shown in Table 2, and shown as bar charts in Fig. 2. The Fragment Library from Specs has the expected distribution of properties given the criteria used for identifying the Library, with the most striking feature being the hard cut-off for solubility. The property distribution is similar for the other Fragment Libraries from Maybridge and Asinex (as would be expected given the strict criteria applied for selection), although the Asinex derived fragments are overall slightly larger compounds with more rotatable bonds, rings, Hbond acceptors and slightly more negatively charged.

Fragment Sets were derived and profiled for each of the six protocols for each Fragment Library from the three suppliers. In this section, the characteristics of the process and the properties of the different Fragment Sets are discussed in detail for one of the suppliers (Specs), with additional comments on any differences seen for the other two suppliers.

The properties of the six different Fragment Sets derived from the Specs Fragment Library are listed in Table 2 and illustrated in Fig. 2. Four main comments can be made about the properties of the different Fragment Sets compared to the parent Fragment Library. First, "Cluster All", "Cluster Fragment" and "Iterative Removal" select smaller compounds, as reflected in the molecular weight and heavy atom count. "Similarity within Cluster" selects compounds with representative properties, whereas the two Substructure protocols select larger compounds with higher ALogP values. Secondly, "Iterative Removal" selects positively charged fragments, whereas the other protocols

Table 2 Physico-chemical properties of Fragment Libraries from the different suppliers and the Fragment Sets resulting from the different protocols for the Specs Fragment Library

Library	Molecular weight (MW)	No. heavy atoms (AC)	Formal charge (FC)	ALogP	No. HBD Acc (HA)
Asinex	216.5 ±33.3	15.3 ±2.3	−0.23 ±0.47	1.11 ±1.03	3.06 ±0.99
Maybridge	203.4 ±30.7	14.3 ±2.2	−0.18 ±0.42	1.36 ±0.89	2.92 ±1.03
Specs	205.4 ±37.7	14.5 ±2.6	−0.11 ±0.42	1.35 ±0.94	2.81 ±1.03
Cluster All	185.40 ±37.8	13.00 ±2.8	−0.07 ±0.38	1.33 ±0.92	2.48 ±1.18
Cluster Fragments	179.52 ±36.9	12.70 ±2.8	0.01 ±0.38	1.33 ±0.86	2.51 ±1.12
Sim within Cluster	201.90 ±38.4	14.27 ±2.8	−0.04 ±0.43	1.48 ±0.92	2.77 ±1.18
Substructure Count	220.69 ±28.2	16.33 ±1.7	0.00 ±0.40	1.95 ±0.62	2.66 ±1.07
Substructure Map	237.74 ±26.3	17.25 ±1.5	−0.01 ±0.22	2.10 ±0.65	2.60 ±0.93
Iterative Removal	174.85 ±31.6	12.22 ±2.2	−0.11 ±0.32	1.38 ±0.96	2.14 ±0.85

Library	No. HBD donors (HD)	No. rot bonds (RB)	PSA	Solubility (LogS)	No. aromatic bonds (ArB)
Asinex	0.76 ±0.76	2.42 ±1.30	57.9 ±14.3	−2.30 ±0.77	7.09 ±3.57
Maybridge	0.75 ±0.75	1.94 ±1.27	56.6 ±14.8	−2.34 ±0.69	7.10 ±3.50
Specs	0.78 ±0.71	2.18 ±1.45	54.1 ±15.3	−2.33 ±0.74	6.90 ±3.31
Cluster All	0.66 ±0.62	1.64 ±1.36	49.54 ±17.0	−2.21 ±0.79	6.70 ±4.13
Cluster Fragments	0.74 ±0.71	1.59 ±1.35	48.92 ±17.6	−2.11 ±0.84	6.54 ±3.92
Sim within Cluster	0.72 ±0.67	1.90 ±1.43	52.35 ±16.1	−2.30 ±0.88	7.51 ±3.98
Substructure Count	0.85 ±0.78	2.21 ±1.11	49.30 ±16.7	−3.01 ±0.26	12.73 ±2.03
Substructure Map	0.68 ±0.81	1.77 ±0.93	47.24 ±17.0	−2.79 ±0.44	8.76 ±3.92
Iterative Removal	0.45 ±0.62	1.54 ±1.26	42.70 ±16.1	−2.17 ±0.77	6.02 ±2.98

Library	No. rings (R)	No. aromatic rings (ArR)	No. ring assemblies (RA)
Asinex	1.93 ±0.63	1.28 ±0.68	1.54 ±0.52
Maybridge	1.84 ±0.73	1.27 ±0.67	1.43 ±0.55
Specs	1.79 ±0.71	1.22 ±0.63	1.41 ±0.52
Cluster All	1.57 ±0.74	1.17 ±0.64	1.23 ±0.48
Cluster Fragments	1.59 ±0.74	1.16 ±0.72	1.27 ±0.51
Iterative Removal	1.91 ±0.73	1.35 ±0.75	1.43 ±0.55
Substructure Count	2.39 ±0.52	2.20 ±0.43	2.04 ±0.33
Substructure Map	2.91 ±0.68	1.51 ±0.71	2.12 ±0.44
Iterative Removal	1.45 ±0.70	1.03 ±0.54	1.22 ±0.45

select some negatively charged ones, giving the (overall small) shift in the charged nature of the Fragment Sets. Also, compounds in the "Iterative Removal" Fragment Set have fewer hydrogen bond donors and acceptors. Thirdly, the "Similarity within Cluster" Fragment Set has the closest property profile to the Fragment Library. Finally, there are surprising differences in the properties of the compounds selected by the two "Substructure" protocols. The Fragment Set produced by the "Count" method has more hydrogen bond donors and acceptors, more rotatable bonds and more aromatic bonds, whereas the Fragment Set produced by the "Map" method has more rings and ring assemblies, a higher ALogP and larger compounds. The Fragment Sets derived by the six Protocols for the other suppliers show the same pattern of differences in characteristics (summarised in Table S4).

Some of these differences in properties can be rationalised by the nature of the protocols used. In "Similarity within Cluster", the compounds are clustered first and then the most representative fragment is selected. Assuming that clustering on FCFP_4 properties effectively clusters on the physicochemical properties, then it is not surprising this Fragment Set has the most similar properties to the Fragment Library. On the other hand, the "Substructure" protocols will select for larger fragments as more substructures will be present in such fragments and thus a higher score

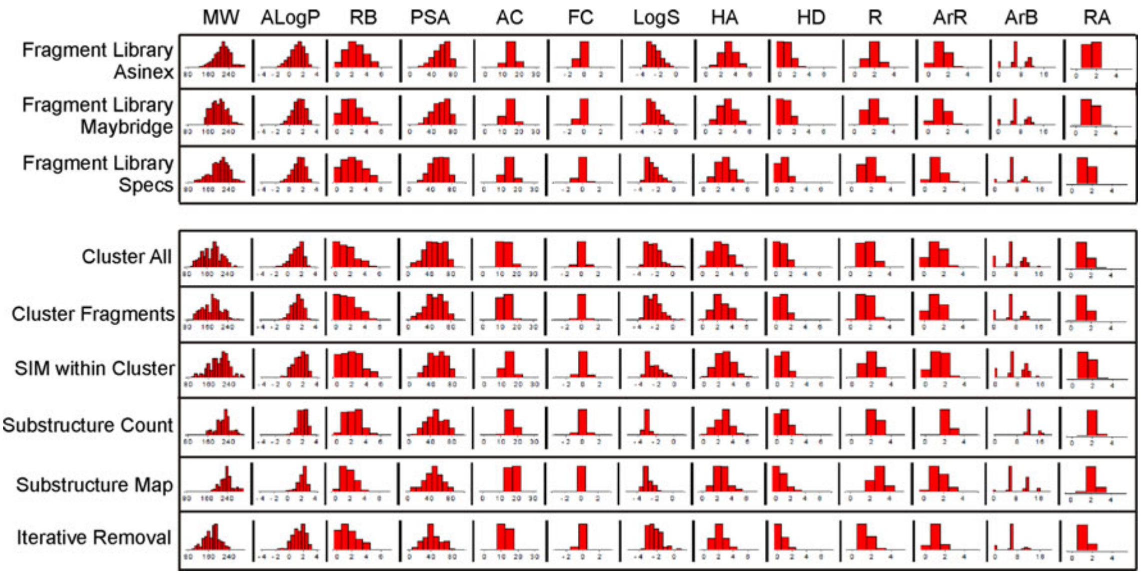

Fig. 2 Properties of Fragment Libraries from three suppliers and the Fragment Sets derived from the Specs Input Library

obtained for the presence of substructures in the Non-Fragments. The differences between "Substructure Count" and "Substructure Map" are due to the way bonds and ring features are counted. For example, if a molecule has 7 more bonds and 1 more ring, then "Count" will give an increased score of 8, whereas "Map" would give an increased score of just 2. The differences seen in the "Iterative Removal" set is probably because the smaller fragments will have a higher Tversky similarity to more Non-Fragments and will thus be selected in the first part of that Protocol. Such smaller fragments will also have a lower number of hydrogen bond donors and acceptors as well. However, the small differences seen in the average formal charge across the Fragment Sets is difficult to explain. It may not be significant, given the small number of compounds with charges that are present in the dataset.

Table 3 provides a profile of the overall characteristics of the Fragment Sets, considering the properties that are

Table 3 Profile of the Fragment Sets generated by the different Protocols from different suppliers

Protocol	Supplier library	SIMILARITY*	NON-SIMILAR*	DRUG_LIKE_1*	DRUG_LIKE_2*	DIVERSITY_1*	DIVERSITY_2*
Cluster All	Asinex	217,156	246,820	0.421	0.307	0.201	6.34
	Maybridge	23,446	28,190	0.373	0.295	0.172	7.62
	Specs	95,103	100,239	0.406	0.309	0.179	8.71
Cluster Fragments	Asinex	218,136	247,350	0.407	0.298	0.184	8.27
	Maybridge	23,866	28,183	0.367	0.291	0.162	9.44
	Specs	91,572	101,211	0.384	0.301	0.159	8.62
SIM within Cluster	Asinex	178,519	249,631	0.367	0.267	0.164	8.40
	Maybridge	18,848	28,385	0.331	0.264	0.145	8.50
	Specs	75,175	102,367	0.343	0.261	0.147	10.28
Substructure Count	Asinex	108,589	249,439	0.390	0.267	0.256	10.03
	Maybridge	11,072	28,446	0.337	0.250	0.200	9.54
	Specs	45,757	102,508	0.367	0.265	0.235	11.42
Substructure Map	Asinex	121,667	250,317	0.367	0.276	0.201	11.42
	Maybridge	11,865	28,462	0.349	0.281	0.191	6.79
	Specs	50,207	102,780	0.373	0.284	0.201	8.07
Iterative Removal	Asinex	229,040	245,541	0.436	0.319	0.198	10.07
	Maybridge	25,198	27,976	0.407	0.312	0.198	8.60
	Specs	99,273	98,749	0.444	0.335	0.193	8.18

* See "Methods" section for description of the calculations

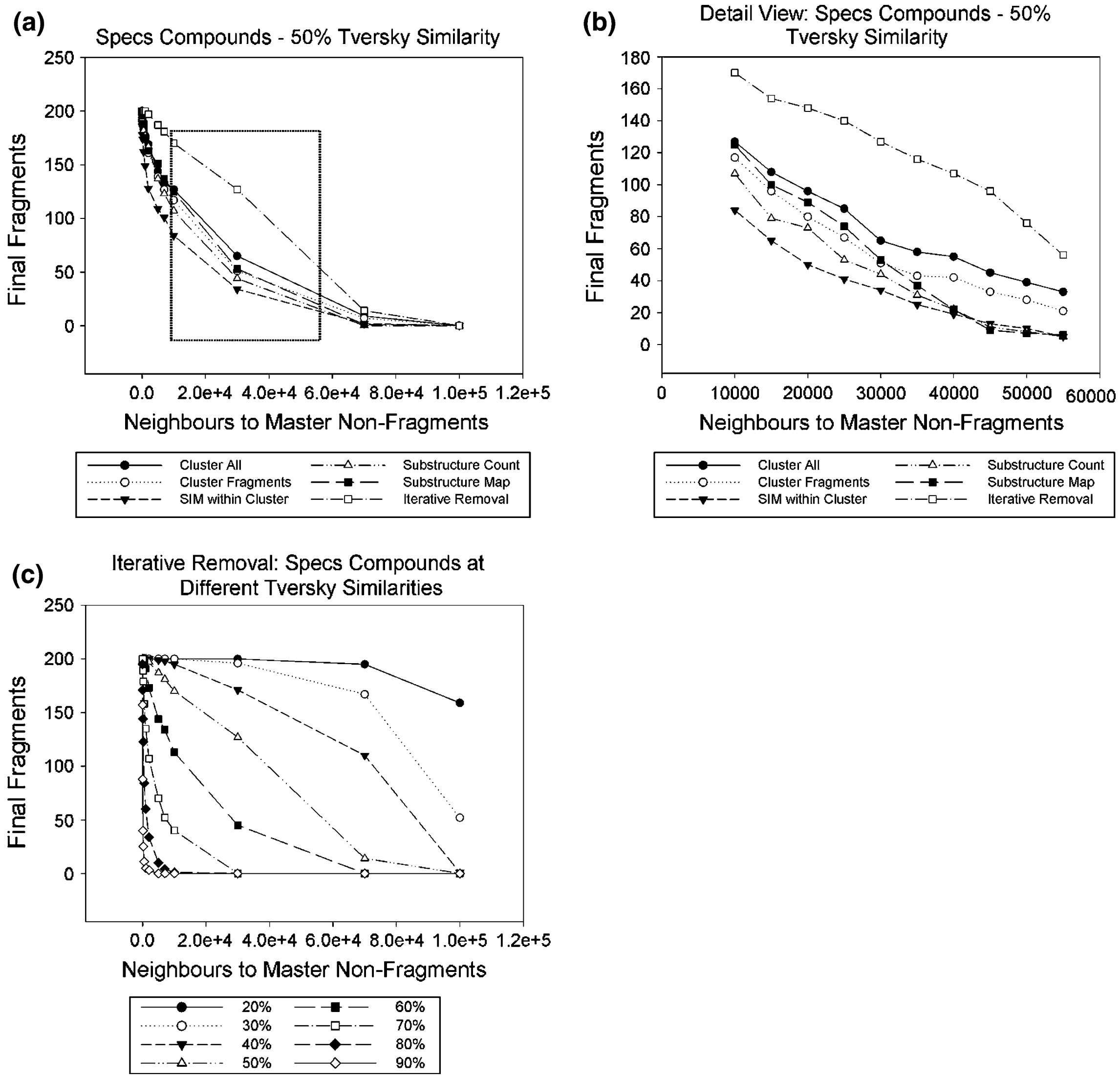

Fig. 3 For Specs derived Fragment Sets—number of NN compounds at 50% Tversky similarity for different Fragment Sets and number of NN compounds for different Tversky similarity values for the Iterative Removal Fragment Set

important for a screening collection. For these properties, the best library would be one that has a high SIMILARITY (and low NON-SIMILARITY) to the Non-Fragment Library, contains a diverse collection of fragments (low DIVERSITY scores), and is DRUG-LIKE. On these criteria, the Iterative Removal is overall the best performing protocol, with the exception of DIVERSITY scores, where it is average. As the SIMILARITY score is the primary aim of the library design, this has been analysed in more detail. Figure 3a plots the number of fragments (*y*-axis) from each of the Fragment Sets that have more than a certain number (*x*-axis) of compounds in the Non-Fragment Library from Specs which are more than 50% similar by Tversky. This metric reinforces the SIMILARITY calculation and shows that the Iterative Removal protocol is the most effective at generating a library that has the greatest coverage of the Non-Fragment Library. Figure 3b shows how this Fragment Set covers the Non-Fragment Library at range of Tversky similarity from 20 to 90%.

Concluding remarks

We have developed and investigated different protocols for selecting Fragment Sets that are representative of a compound library. The “Iterative Removal” protocol generates the best Fragment Set, judged on the similarity to the Non-Fragment Library and the overall characteristics of the Set. These protocols are relatively straightforward to implement and could be used to select fragments for screening with an increased probability that nearest neighbour compounds will be available for subsequent fragment evolution. Such

Fragment Sets have been acquired and used in various screening campaigns in our institution and results will be described in due course.

We have described and made available the protocols as the approach could prove useful to others for the design of fragment screening libraries that represent commercially available libraries, the compounds available in a proprietary in-house collection or a virtual library of compounds that could be rapidly synthesised given available resources.

Acknowledgments We thank Accelrys for providing access to software and Eddy Van de Water of Accelrys for specific technical support. We are grateful to Hans Briem and Judith Günther from Bayer and Steve Roughley and James Davidson of Vernalis for helpful discussions. The work was supported by a grant from the Biotechnology and Biological Sciences Research Council, JL by a fellowship from the Swedish Pharmaceutical Society and MNS was partially supported by the Wild fund.

References

1. Shuker SB, Hajduk PJ, Meadows RP, Fesik SW (1996) Discovering high-affinity ligands for proteins: SAR by NMR. Science 274(5292):1531–1534
2. Hajduk PJ (2006) SAR by NMR: putting the pieces together. Mol Interv 6(5):266–272. doi:10.1124/mi.6.5.8
3. Fischer M, Hubbard RE (2009) Fragment-based ligand discovery. Mol Interv 9(1):22–30. doi:10.1124/mi.9.1.7
4. Schulz MN, Hubbard RE (2009) Recent progress in fragment-based lead discovery. Curr Opin Pharmacol 9(5):615–621. doi:10.1016/j.coph.2009.04.009
5. Congreve M, Chessari G, Tisi D, Woodhead AJ (2008) Recent developments in fragment-based drug discovery. J Med Chem 51(13):3661–3680. doi:10.1021/jm8000373
6. Hopkins AL, Groom CR, Alex A (2004) Ligand efficiency: a useful metric for lead selection. Drug Discov Today 9(10):430–431. doi:10.1016/S1359-6446(04)03069-7
7. Lepre CA (2011) Practical aspects of NMR-based fragment screening. Methods Enzymol 493:219–239. doi:10.1016/B978-0-12-381274-2.00009-1
8. Murray CW, Blundell TL (2010) Structural biology in fragment-based drug design. Curr Opin Struct Biol 20(4):497–507. doi:10.1016/j.sbi.2010.04.003
9. Giannetti AM (2011) From experimental design to validated hits a comprehensive walk-through of fragment lead identification using surface plasmon resonance. Methods Enzymol 493:169–218. doi:10.1016/B978-0-12-381274-2.00008-X
10. Kranz JK, Schalk-Hihi C (2011) Protein thermal shifts to identify low molecular weight fragments. Methods Enzymol 493:277–298. doi:10.1016/B978-0-12-381274-2.00011-X
11. Hubbard RE, Murray JB (2011) Experiences in fragment-based lead discovery. Methods Enzymol 493:509–531. doi:10.1016/B978-0-12-381274-2.00020-0
12. Erlanson DA (2006) Fragment-based lead discovery: a chemical update. Curr Opin Biotechnol 17(6):643–652. doi:10.1016/j.copbio.2006.10.007
13. Barker JJ, Barker O, Courtney SM, Gardiner M, Hesterkamp T, Ichihara O, Mather O, Montalbetti CA, Muller A, Varasi M, Whittaker M, Yarnold CJ (2010) Discovery of a novel Hsp90 inhibitor by fragment linking. ChemMedChem 5(10):1697–1700. doi:10.1002/cmdc.201000219
14. Brough PA, Barril X, Borgognoni J, Chene P, Davies NG, Davis B, Drysdale MJ, Dymock B, Eccles SA, Garcia-Echeverria C, Fromont C, Hayes A, Hubbard RE, Jordan AM, Jensen MR, Massey A, Merrett A, Padfield A, Parsons R, Radimerski T, Raynaud FI, Robertson A, Roughley SD, Schoepfer J, Simmonite H, Sharp SY, Surgenor A, Valenti M, Walls S, Webb P, Wood M, Workman P, Wright L (2009) Combining hit identification strategies: fragment-based and in silico approaches to orally active 2-aminothieno[2,3-d]pyrimidine inhibitors of the Hsp90 molecular chaperone. J Med Chem 52(15):4794–4809. doi:10.1021/jm900357y
15. Hubbard RE (2008) Fragment approaches in structure-based drug discovery. J Synchrotron Radiat 15(Pt 3):227–230. doi:10.1107/S090904950705666X
16. Mochalkin I, Miller JR, Narasimhan L, Thanabal V, Erdman P, Cox PB, Prasad JV, Lightle S, Huband MD, Stover CK (2009) Discovery of antibacterial biotin carboxylase inhibitors by virtual screening and fragment-based approaches. ACS Chem Biol 4(6):473–483. doi:10.1021/cb9000102
17. Davies TG, Woodhead SJ, Collins I (2009) Fragment-based discovery of inhibitors of protein kinase B. Curr Top Med Chem 9(18):1705–1717
18. Orita M, Ohno K, Warizaya M, Amano Y, Niimi T (2011) Lead generation and examples opinion regarding how to follow up hits. Methods Enzymol 493:383–419. doi:10.1016/B978-0-12-381274-2.00015-7
19. Brough PA, Aherne W, Barril X, Borgognoni J, Boxall K, Cansfield JE, Cheung KM, Collins I, Davies NG, Drysdale MJ, Dymock B, Eccles SA, Finch H, Fink A, Hayes A, Howes R, Hubbard RE, James K, Jordan AM, Lockie A, Martins V, Massey A, Matthews TP, McDonald E, Northfield CJ, Pearl LH, Prodromou C, Ray S, Raynaud FI, Roughley SD, Sharp SY, Surgenor A, Walmsley DL, Webb P, Wood M, Workman P, Wright L (2008) 4,5-diarylisoxazole Hsp90 chaperone inhibitors: potential therapeutic agents for the treatment of cancer. J Med Chem 51(2):196–218. doi:10.1021/jm701018h
20. Crisman TJ, Bender A, Milik M, Jenkins JL, Scheiber J, Sukuru SC, Fejzo J, Hommel U, Davies JW, Glick M (2008) “Virtual fragment linking”: an approach to identify potent binders from low affinity fragment hits. J Med Chem 51(8):2481–2491. doi:10.1021/jm701314u
21. Lau W, Withka J, Hepworth D, Magee T, Du Y, Bakken G, Miller M, Hendsch Z, Thanabal V, Kolodziej S, Xing L, Hu Q, Narasimhan L, Love R, Charlton M, Hughes S, van Hoorn W, Mills J (2011) Design of a multi-purpose fragment screening library using molecular complexity and orthogonal diversity metrics. J Comput Aided Mol Des 1–16. doi:10.1007/s10822-011-9434-0
22. Baurin N, Aboul-Ela F, Barril X, Davis B, Drysdale M, Dymock B, Finch H, Fromont C, Richardson C, Simmonite H, Hubbard RE (2004) Design and characterization of libraries of molecular fragments for use in NMR screening against protein targets. J Chem Inf Comput Sci 44(6):2157–2166. doi:10.1021/ci049806z
23. Hubbard RE, Chen I, Davis B (2007) Informatics and modeling challenges in fragment-based drug discovery. Curr Opin Drug Discov Devel 10(3):289–297
24. Chen IJ, Hubbard RE (2009) Lessons for fragment library design: analysis of output from multiple screening campaigns. J Comput Aided Mol Des. doi:10.1007/s10822-009-9280-5
25. Blomberg N, Cosgrove D, Kenny P, Kolmodin K (2009) Design of compound libraries for fragment screening. J Comput Aided Mol Des 23(8):513–525. doi:10.1007/s10822-009-9264-5
26. Gianti E, Sartori L (2008) Identification and selection of “privileged fragments” suitable for primary screening. J Chem Inf Model 48(11):2129–2139. doi:10.1021/ci800219h

27. Venhorst J, Núñez S, Kruse CG (2010) Design of a high fragment efficiency library by molecular graph theory. ACS Med Chem Lett 1(9):499–503. doi:10.1021/ml100163s
28. Tounge BA, Parker MH (2011) Designing a diverse high-quality library for crystallography-based FBDD screening. Methods Enzymol 493:3–20. doi:10.1016/B978-0-12-381274-2.00001-7
29. Prakesch M, Denisov AY, Naim M, Gehring K, Arya P (2008) The discovery of small molecule chemical probes of Bcl-X(L) and Mcl-1. Bioorg Med Chem 16(15):7443–7449. doi:10.1016/j.bmc.2008.06.023
30. Dalvit C, Mongelli N, Papeo G, Giordano P, Veronesi M, Moskau D, Kummerle R (2005) Sensitivity improvement in 19F NMR-based screening experiments: theoretical considerations and experimental applications. J Am Chem Soc 127(38):13380–13385. doi:10.1021/ja0542385
31. Blaney J, Nienaber V, Burley SK (2006) Fragment-based lead discovery and optimization using X-ray crystallography, computational chemistry and high-throughput organic synthesis. In: Jahnke W, Erlanson DA (eds) Fragment-based approaches in drug discovery, Wiley, Weinheim, Germany
32. Irwin JJ, Shoichet BK (2005) ZINC—a free database of commercially available compounds for virtual screening. J Chem Inf Model 45(1):177–182. doi:10.1021/ci049714+
33. Tversky A (1977) Features of similarity. Psychol Rev 84(4):327–352. doi:10.1037/0033-295x.84.4.327