Molecular Complexity and Its Impact on the Probability of Finding Leads for Drug Discovery

Michael M. Hann,* Andrew R. Leach, and Gavin Harper

Computational Chemistry and Informatics Unit, GlaxoSmithKline Medicines Research Centre, Gunnels Wood Road, Stevenage, SG1 2NY, England

Received November 24, 2000

Using a simple model of ligand−receptor interactions, the interactions between ligands and receptors of varying complexities are studied and the probabilities of binding calculated. It is observed that as the systems become more complex the chance of observing a useful interaction for a randomly chosen ligand falls dramatically. The implications of this for the design of combinatorial libraries is explored. A large set of drug leads and optimized compounds is profiled using several different properties relevant to molecular recognition. The changes observed for these properties during the drug optimization phase support the hypothesis that less complex molecules are more common starting points for the discovery of drugs. An extreme example of the use of simple molecules for directed screening against thrombin is provided.

INTRODUCTION

In recent years enormous effort has been expended trying to increase the success rate of the drug discovery process. One aspect of this has been the use of technologies for synthesizing and screening large libraries of molecules. Despite some successes, it is apparent that the high throughput synthesis and screening paradigm has not delivered the results that were initially anticipated.[1,2] Three probable and inter-related reasons for this lack of success are (1) the immaturity of the technology, (2) the inability to make the right types of molecules with the technology, and (3) a lack of understanding of what the right types of molecules to make actually are. Much has been written about the first two of these aspects, but there is less discussion about issues related to the third aspect. Although concepts of drug likeness have been widely investigated, these are invariably concerned with identifying properties of molecules which are needed to ensure that proposed molecules are drug-like.[3−5] In this paper we discuss a different aspect of this problem concerning the relationship between the probability of discovering lead molecules from screening and their molecular complexity. We present both an analysis of existing data and a theoretical model that helps to explore the issues. Teague and colleagues have discussed related issues to this in a recent paper but did not consider the theoretical analysis presented here.[6]

COMPLEXITY OF MOLECULES AND MODES OF BINDING

The binding of a small molecule to a defined binding site is a very complex balance of many components, collectively giving rise to the free energy of binding. This has both entropic and enthalpic components. In addition there has to be a kinetically accessible pathway by which the bound state can be achieved, which also has entropic and enthalpic components.[7,9] The processes that lead to successful binding are often referred to as *molecular recognition*.

The aspect of molecular recognition that relates to the work described here concerns the probability that any one molecule's features are compatible with those of a designated binding site. For real molecular systems calculation of the overall energetics of all aspects of binding requires a prohibitively large amount of computation, and it is common practice to use simplifying models to help predict ligand binding energetics. For example, fast scoring functions to represent free energy estimates,[10] extended QSAR methods such as 4D-QSAR,[11] and pharmacophore and database search methods.[12]

In the same way we have chosen to use a very simple but still functional model to represent probabilistic aspects of molecular recognition and drug discovery which we believe have not previously been discussed.

Effective molecular recognition is essentially the matching of surface properties of a ligand molecule with its binding pocket through complementarity of shape and electronic properties such as charge and hydrophobicitiy. These can be considered in their simplest form to be localized recognition elements that are highly detrimental to binding if they are not correctly matched but beneficial if correct. A very simple model of molecular recognition can be represented by a pattern of +'s and −'s that represent the features of a binding site and another pattern that represents the ligand molecule. We count as a successful recognition event a ligand in which all the + features match a binding site − feature (or visa versa). These features represent any aspect of a molecule (shape, dipolarity and higher moments, hydrophobicity, etc.) that need to be matched to the binding site and are not merely indicative of formal charges or partial charges such as H-bond interactions. The model allows the binding site to contain variable number of features by expanding the length of the pattern representing it. Possible ligands are then restricted in pattern length to be equal to or less than that of the binding site to mimic the fact that most binding sites

* Corresponding author phone: ++44 1438 763392; fax: ++44 1438 764918; e-mail: mmh1203@gsk.com.

10.1021/ci000403i CCC: $20.00
Published on Web 04/27/2001

Feature Position	1	2	3	4	5	6	7	8	9
Receptor features	-	-	+	-	+	-	-	+	-
Ligand mode 1	+	+	-						
Ligand mode 2						+	+	-	
(reverse mode 3)					-	+	+		
(end wrap mode 4)	+							-	+

Figure 1. Example of ligand/receptor matching possibilities for a random receptor of complexity 9 and a ligand of complexity 3.

have a size and shape that will limit the size of a ligand. While it is true that parts of a ligand may protrude from the binding site and therefore not contribute to the binding, we are only interested here in recognition within the site. Our model does not incorporate flexibility in either the ligand or the binding side; such flexibility could, to a certain extent, remove bad matches and optimize good ones. However, we know that freezing out of motion due to rotation about rotatable bonds when a molecule binds has its own entropic penalties, and it is generally beneficial to limit the amount of flexibility in ligands.[8,9] As mentioned earlier, our model is highly simplified and abstract from real ligand molecules and binding sites as it is designed to illustrate a particular issue. It should be understood that any molecular properties can in principle be represented as a pattern of changing values surrounding a molecule. The one-dimensional pattern used here is a linear representation of much more complex patterns that exist in real molecules.

To illustrate this model, examine the representation of a binding site with nine features (chosen at random from +'s and −'s) shown in Figure 1 and consider how many different ways a ligand with complexity of three features represented as + + − can match. In this example, there are two ways the ligand can match, as shown. (In our model we have chosen to exclude reverse binding such as mode 3 and end wrap matches such as mode 4, although the statistics could be easily adjusted to account for this.)

With this type of model it is possible to explore how the complexity of ligands (as indicated by varying length and pattern of features) effects its chance of matching a binding site of given complexity. We can use this model to calculate the probability that a randomly chosen "molecule" (such as a single member of a chemical library or screening collection) might exactly fit a given binding site. We can then consider the effect on the probability of changing the information content of the ligand's complexity in relation to the degree of complexity present in the binding site. In our model one mismatch is considered sufficient to totally obviate binding. This may seem a harsh criterion but in our experience of studying structure−activity relationships, it is not uncommon for a very minor change to cause a dramatic decrease in binding to a specified target. Mismatches could be incorporated into the model; however, they will not effect the overall conclusions but soften the severity of the model. A perfect match represents the best (i.e. maximal energy) binding that a ligand can achieve for its given level of complexity. The questions that we wish to explore are "what are the chances of finding matching molecules as their complexity increases" and "how does complexity effect the uniqueness of binding modes".

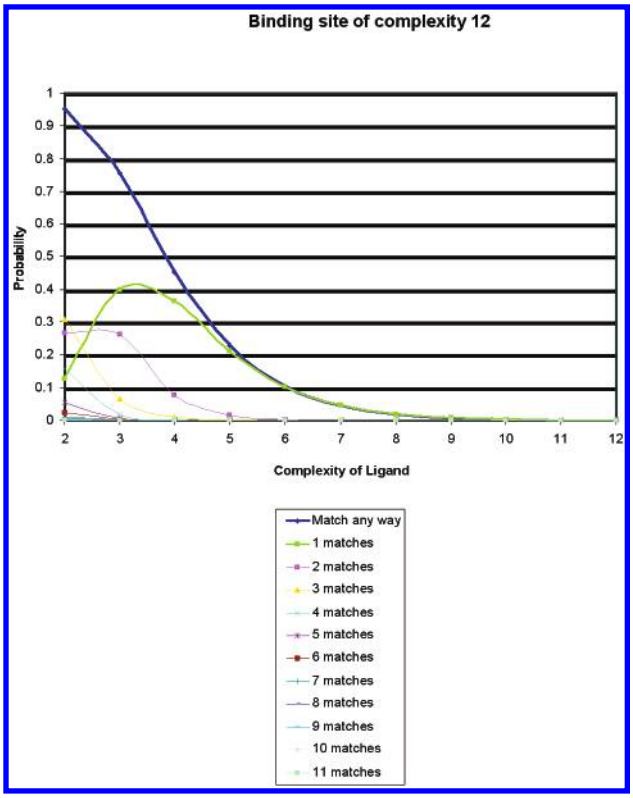

Figure 2. Probabilities of ligands of varying complexity matching a binding site of complexity 12.

Using this model system, we have exhaustively computed for various levels of complexity of binding site and ligands the probability that a randomly selected ligand of a given complexity matches a binding site and calculations were performed by enumerating all possible ligand ± patterns of varying lengths and comparing them to enumerated ± binding sites. We also compute, as was illustrated in Figure 1 for a receptor of complexity 9 and ligand of complexity 3, the number of different positions in which the ligand can match. The data for simulations of binding sites up to and including 12 features are provided in Table 1, while graphical representations of the data for the 12 feature situation are shown in Figures 2 and 3.

In Figure 2, it can be seen that the probability of finding any type of match decays exponentially as the complexity of the ligand increases (dark blue). This is because there rapidly become more ways of obtaining a mismatch than a match as the complexity of the ligand increases. It can also be seen (green line) that the chance of finding a ligand that matches only one way (i.e. unique binding mode) in a random choice of ligand actually peaks at a ligand complexity of three. For less complex ligands it is more likely that greater than one match (e.g. 2, 3, or 4 matches) will be found. For longer molecules the chance of finding any hits falls off dramatically. This probability of there being a unique binding mode is related to the unfrustrated energy landscape discussed by Rejto and Verkhivker.[13] They highlight the virtue of such a unique binding mode in docking studies of small molecular "anchors" as the key to detecting those that will have favorable kinetic access to the site.

In Figure 3 (which we refer to as the "success landscape") we have introduced a new curve to represent the probability of actually **measuring** a binding interaction that has been

Table 1. Simulation Results[a]

M = complexity of molecule	probability of match anyway	0 ways	1 way	2 ways	3 ways	4 ways	5 ways	6 ways	7 ways	8 ways	9 ways	10 ways	11 ways
						$N = 2$[b]							
$M = 2$	0.25	0.75	0.25										
						$N = 3$[b]							
$M = 2$	0.438	0.562	0.375	0.062									
$M = 3$	0.125	0.875	0.125	0									
						$N = 4$[b]							
$M = 2$	0.594	0.406	0.469	0.094	0.031								
$M = 3$	0.234	0.766	0.219	0.016	0								
$M = 4$	0.062	0.938	0.062	0	0								
						$N = 5$[b]							
$M = 2$	0.703	0.297	0.469	0.188	0.031	0.016							
$M = 3$	0.336	0.664	0.305	0.023	0.008	0							
$M = 4$	0.121	0.879	0.117	0.004	0	0							
$M = 5$	0.031	0.969	0.031	0	0	0							
						$N = 6$[b]							
$M = 2$	0.781	0.219	0.43	0.266	0.062	0.016	0.008						
$M = 3$	0.426	0.574	0.367	0.047	0.008	0.004	0						
$M = 4$	0.178	0.822	0.17	0.006	0.002	0	0						
$M = 5$	0.062	0.938	0.061	0.001	0	0	0						
$M = 6$	0.016	0.984	0.016	0	0	0	0						
						$N = 7$[b]							
$M = 2$	0.836	0.164	0.367	0.332	0.094	0.031	0.008	0.004					
$M = 3$	0.504	0.496	0.408	0.078	0.012	0.004	0.002	0					
$M = 4$	0.231	0.769	0.217	0.012	0.002	0.001	0	0					
$M = 5$	0.091	0.909	0.089	0.001	5e-04	0	0	0					
$M = 6$	0.031	0.969	0.031	2e-04	0	0	0	0					
$M = 7$	0.008	0.992	0.008	0	0	0	0	0					
						$N = 8$[b]							
$M = 2$	0.875	0.125	0.303	0.363	0.146	0.039	0.018	0.004	0.002				
$M = 3$	0.571	0.429	0.432	0.112	0.02	0.005	0.002	0.001	0				
$M = 4$	0.282	0.718	0.258	0.02	0.003	0.001	5e-04	0	0				
$M = 5$	0.12	0.88	0.117	0.003	5e-04	2e-04	0	0	0				
$M = 6$	0.046	0.954	0.046	4e-04	1e-04	0	0	0	0				
$M = 7$	0.016	0.984	0.016	6e-05	0	0	0	0	0				
$M = 8$	0.004	0.996	0.004	0	0	0	0	0	0				
						$N = 9$[b]							
$M = 2$	0.903	0.097	0.244	0.365	0.201	0.059	0.021	0.01	0.002	0.001			
$M = 3$	0.629	0.371	0.439	0.151	0.027	0.008	0.002	0.001	5e-04	0			
$M = 4$	0.329	0.671	0.292	0.031	0.004	0.001	5e-04	2e-04	0	0			
$M = 5$	0.149	0.851	0.143	0.005	0.001	2e-04	1e-04	0	0	0			
$M = 6$	0.061	0.939	0.06	0.001	1e-04	6e-05	0	0	0	0			
$M = 7$	0.023	0.977	0.023	9e-05	3e-05	0	0	0	0	0			
$M = 8$	0.008	0.992	0.008	2e-05	0	0	0	0	0	0			
$M = 9$	0.002	0.998	0.002	0	0	0	0	0	0	0			
						$N = 10$[b]							
$M = 2$	0.924	0.076	0.195	0.343	0.252	0.083	0.031	0.012	0.005	0.001	5e-04		
$M = 3$	0.678	0.322	0.434	0.192	0.037	0.01	0.004	0.001	5e-04	2e-04	0		
$M = 4$	0.373	0.627	0.321	0.044	0.005	0.002	0.001	2e-04	1e-04	0	0		
$M = 5$	0.176	0.824	0.167	0.008	0.001	3e-04	1e-04	6e-05	0	0	0		
$M = 6$	0.076	0.924	0.075	0.001	2e-04	6e-05	3e-05	0	0	0	0		
$M = 7$	0.031	0.969	0.031	2e-04	3e-05	2e-05	0	0	0	0	0		
$M = 8$	0.012	0.988	0.012	2e-05	8e-06	0	0	0	0	0	0		
$M = 9$	0.004	0.996	0.004	4e-06	0	0	0	0	0	0	0		
$M = 10$	0.001	0.999	0.001	0	0	0	0	0	0	0	0		
						$N = 11$[b]							
$M = 2$	0.94	0.06	0.156	0.308	0.289	0.119	0.04	0.018	0.007	0.003	5e-04	2e-04	
$M = 3$	0.72	0.28	0.419	0.229	0.05	0.014	0.005	0.002	0.001	2e-04	1e-04	0	
$M = 4$	0.415	0.585	0.345	0.06	0.007	0.002	0.001	3e-04	1e-04	6e-05	0	0	
$M = 5$	0.203	0.797	0.189	0.011	0.001	4e-04	2e-04	6e-05	3e-05	0	0	0	
$M = 6$	0.091	0.909	0.089	0.002	2e-04	8e-05	3e-05	2e-05	0	0	0	0	
$M = 7$	0.039	0.961	0.038	3e-04	5e-05	2e-05	8e-06	0	0	0	0	0	
$M = 8$	0.016	0.984	0.015	5e-05	8e-06	4e-06	0	0	0	0	0	0	
$M = 9$	0.006	0.994	0.006	6e-06	2e-06	0	0	0	0	0	0	0	
$M = 10$	0.002	0.998	0.002	1e-06	0	0	0	0	0	0	0	0	
$M = 11$	5e-04	ca. 1	5e-04	0	0	0	0	0	0	0	0	0	

Table 1 (Continued)

M = complexity of molecule	probability of match anyway	0 ways	1 way	2 ways	3 ways	4 ways	5 ways	6 ways	7 ways	8 ways	9 ways	10 ways	11 ways
						$N = 12$[b]							
$M = 2$	0.952	0.048	0.126	0.266	0.308	0.159	0.054	0.023	0.011	0.004	0.002	2e-04	1e-04
$M = 3$	0.756	0.244	0.399	0.263	0.066	0.018	0.006	0.003	0.001	3e-04	1e-04	6e-05	0
$M = 4$	0.453	0.547	0.364	0.076	0.009	0.002	0.001	4e-04	2e-04	6e-05	3e-05	0	0
$M = 5$	0.228	0.772	0.21	0.016	0.002	5e-04	2e-04	8e-05	3e-05	2e-05	0	0	0
$M = 6$	0.105	0.895	0.102	0.003	3e-04	1e-04	4e-05	2e-05	8e-06	0	0	0	0
$M = 7$	0.046	0.954	0.046	5e-04	6e-05	2e-05	8e-06	4e-06	0	0	0	0	0
$M = 8$	0.019	0.981	0.019	8e-05	1e-05	4e-06	2e-06	0	0	0	0	0	0
$M = 9$	0.008	0.992	0.008	1e-05	2e-06	1e-06	0	0	0	0	0	0	0
$M = 10$	0.003	0.997	0.003	1e-06	5e-07	0	0	0	0	0	0	0	0
$M = 11$	0.001	0.999	0.001	2e-07	0	0	0	0	0	0	0	0	0
$M = 12$	2e-04	ca. 1	2e-04	0	0	0	0	0	0	0	0	0	0

[a] This table gives the probability of a randomly selected "molecule" of complexity M with features + or − matching a "binding site" of complexity N with randomly chosen features + or −. The columns refer to the probabilities of matching k ways. [b] N = complexity of binding site.

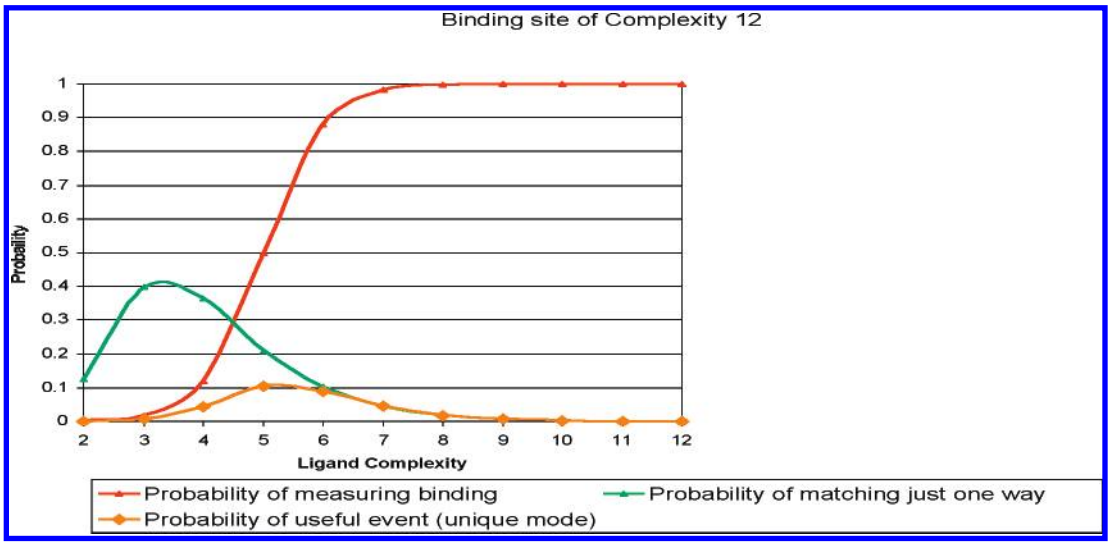

Figure 3. The success landscape.

calculated and plotted (red). This term was calculated by taking the ligand complexity (i.e. x-axis) to represent the number of possible effective interactions that a completely matching ligand can make and then using a hyperbolic function to convert this to the form shown.[14] This form of function was chosen so that small ligands of low complexity have a low probability of exhibiting **measurable** binding, while larger more complex ligands reach plateau value. This reflects the typical nature of real biophysical detection methods where a maximum is reached. It also represents the type of triggering effect that is seen with many biosystems where a threshold is passed, and then no higher effect is registered.

We have then calculated and plotted (the orange curve) the product of this detection probability (red curve) and the probability of finding a unique match for that complexity of ligand (green curve). This product we refer to as the probability of finding a "useful event" for a given molecule.

$$\text{P(useful event)} = \text{P(measure binding)} \times \text{P(ligand matches)}$$

A "useful event" is therefore one that is both measurable and unique in its mode of binding but also has a significant probability of being found, bearing in mind the complexity issue. Thus using this model and the parameters as described, there is clearly a peak at ligand complexity = 5. For ligands of greater complexities there is a diminishing chance of a random ligand having any chance of showing a useful event despite the fact that if it did bind it would have a very strong probability of being measurable. Similarly for complexities less than 4 there is a diminishing chance of a useful event. This is because multiple binding modes begin to dominate for less complex ligands (equivalent to 2 or greater binding modes in Figure 1); the low level of measured affinity of any such ligand will further confound this. It should be noted that the two probabilities that have been multiplied together are independent of each other. Thus one represents the probability of a randomly chosen ligand matching a given binding site,while the other represents the chance that an exactly matched ligand will express sufficient binding to be detected. Also in our model there is no account taken of the influence of factors such as induced flexibility and charge changes (i.e. the binary nature of the patterns are fixed and are not influenced by the pattern found in the binding site). In real systems this can happen and will again influence the ultimate binding mode and energetics, but the characteristics of our model will still be dominant.

While the simplicity of the model and the form of the chosen energy probability can be debated, it is inevitable

that when two probability distributions that have opposing trends are multiplied together the result will have a peak in the middle. In the real world of screening the position, intensity and width of this peak will vary for different receptors and ligand types. Thus the fact that in our severe model the maximum occurs at complexity five should not be used to infer that real molecules with five interaction points (pharmacophores) is the optimal. If, for instance, we introduced into our model the concept of some mismatches being allowed, then the probability of unique binding curve would shift to the right. The result of multiplying this by the energy probability curve would again yield a curve with a maximum in it, but this would be shifted to higher complexity. An alternative situation may be that the measurable energy probability curve may be shifted further to the right. This would then cause the useful event probability curve to be very flat indicating an exceptionally low probability of finding a useful inhibitor. This scenario is probably the situation with looking for small molecule inhibitors of protein protein interactions, which have proved very difficult to find. Success is usually only possible with Phage techniques that provide much larger molecules (which are able to provide sufficient binding because of their size) but at the same time, sample sufficient combinations because of the methodology, used.[15]

Based on the above analysis, our hypothesis is that there is an optimal complexity of molecules that should be considered when screening collections of molecules. Libraries containing very complex molecules have a low chance of individual molecules binding. It is better to start with less complex molecules to get onto the "success landscape" and to increase the potency by increasing the complexity needed.

The model we have used here has the simplest possible (i.e. ±) recognition event. More sophisticated models could be developed which more realistically represent the true 3D nature of property gradients. However, the underlying principles explored with the 1D pattern model will be present in all more sophisticated models. Rather than develop these other models we have chosen to consider data from real-life examples of lead discovery to substantiate the hypotheses derived from the simple model. The purpose of our model is to draw attention to the existence and consequences of the maximum in the "useful event" curve and not to over interpret this model's detail.

THE DIFFERENCE BETWEEN LEADS AND DRUGS

If this theoretical analysis shows the effect of increasing ligand complexity on the chance of observing a useful event, how does this relate to the problem of real molecules in real screens? It is clear that large libraries of synthetic compounds have not had the success rate for finding lead molecules that was originally anticipated, and this is often explained as being due to the molecules made by libraries technologies not being sufficiently drug-like.[2] Many laboratories have now introduced design concepts so the compounds and libraries synthesized made have more drug-like properties. However, based on the above analysis, this may not be sufficient, and we may still be working too far on the right-hand side of the success landscape maximum in Figure 3. The discovery of most drugs invariably incorporates an optimization process whereby activity is adjusted (usually increased) from a starting lead, as other properties (e.g. pharmacokinetic) are enhanced. However, if the complexity of the molecules being screened is already of the order of known drugs, then our options will be limited during the optimization phase. Thus by having too much functionality in library molecules we both decrease the chance of finding a hit molecule in the first place, and then if a hit is found, we narrow the opportunity to, for instance, increase MW as a tool to increase potency. In other words, we are expecting too much, too soon in the way of optimal properties for our screening molecules. Only by working to the left (or at the maximum) of the success landscape can we ensure that we have the best opportunity to identify binding in the first place.

In their analysis of this problem, Teague et al. discussed a small set (18 compounds) of leads and drugs that they had collated. However, in a book entitled "Drug Prototypes and their Exploitation", Walter Sneader extensively reviews the development history of many drugs developed up to the middle of 1990s.[16] He identifies, where possible, the "drug prototype" which is the starting point for the final drug compound. We have developed a Daylight database[17] in which the drug prototypes (i.e. leads) and final drug molecules are linked and can be searched. We ignored those drugs where there is no clear origin assigned and those that are based on heavy metals or other nonclassical drug type. We also removed those leads that were considered to be of microbial or plant origins (about 15% of examples).[18] After further removing counterion salts, the remaining 470 pairs of structures were then profiled using our ADEPT program.[19] This was done so as to be able to compare properties of the lead and drug molecules. In addition the profile of the World Drug Index (ca. 30 000 structures[20]) has been calculated to check that the drug histories used from Sneader's book are representative of a larger set of drug like molecules. The ADEPT profiles for MW, ClogP,[21] number of aromatic rings,[22] Andrew's binding energy,[23] number of bits,[24] CMR,[25] and number of heavy atoms for these three sets of molecules are shown in Figure 4.

It can be seen that the Sneader drugs (green) are indeed close in their profile to the WDI drugs (red). Moreover, the Sneader leads (blue) have profiles for which the distribution is left shifted when compared to their resulting drugs. The mean property value of the Sneader lead set and the absolute and percentage change on going from these leads to the drugs for a range of properties are shown in Table 2. The data indicates that, on average, drug leads have lower MW, lower ClogP, fewer aromatic rings, fewer hydrogen bond acceptors, and lower Andrew's binding energy than their corresponding drugs. The largest percentage change is in the Andrew's binding energy function. This complex function is a summation of many terms related to molecular recognition which may explain why it is particularly good at differentiating leads from drugs. As noted by Teague et al. a significant increase in ClogP is often associated with the lead to drug process reflecting the opportunity to gain potency through localized hydrophobic effects at the receptor and also for tuning the bulk properties for optimal pharmacokinetics.[6] The number of bits set in the Daylight 2D structure representation is an indication of the internal bond complexity of a molecule. That this increases is also consistent with molecules becoming more complex. Interestingly, at least for the set studied here, the changes in hydrogen bond donor numbers are not significant.

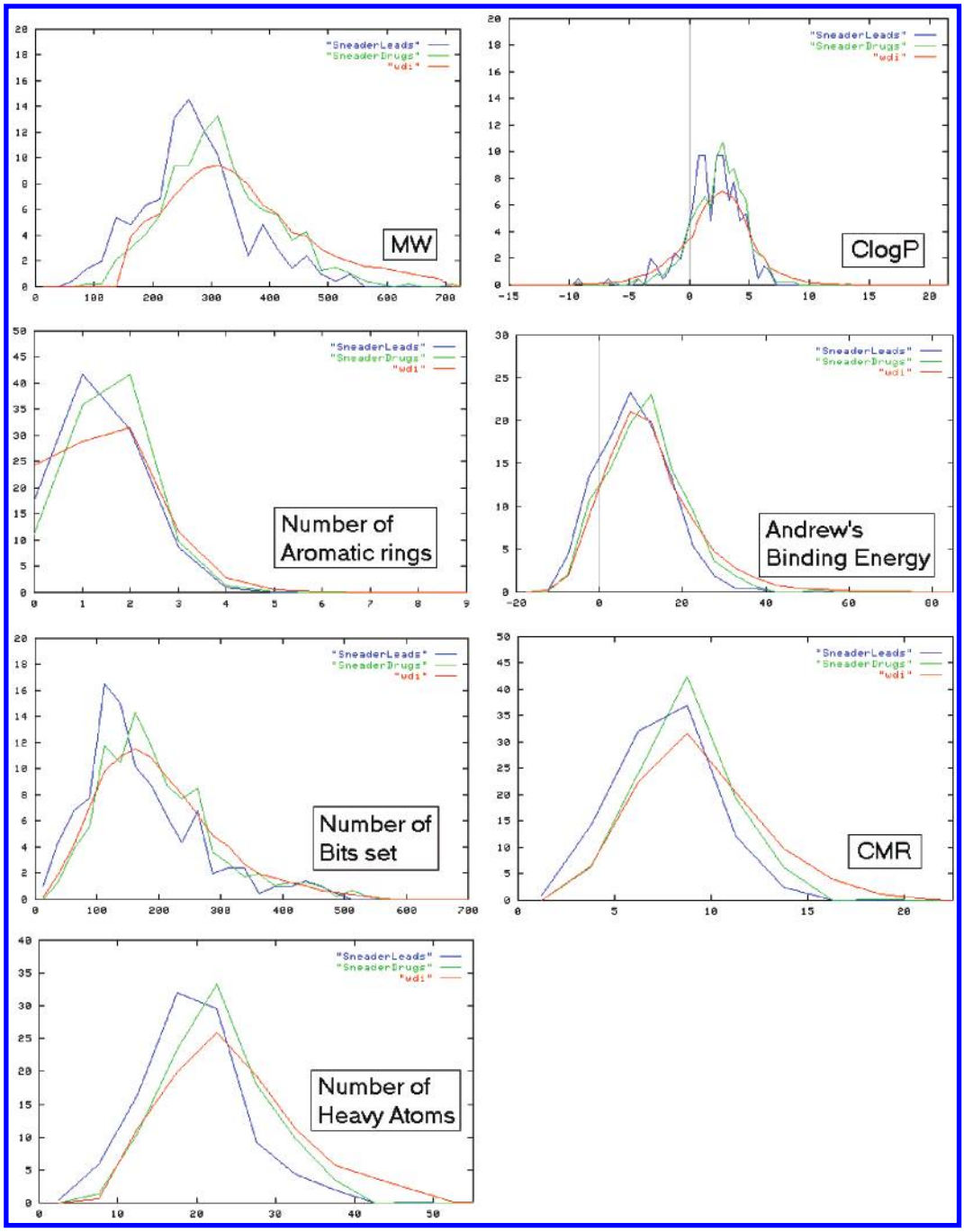

Figure 4. ADEPT plots comparing Sneader leads (blue), Sneader drugs (green), and WDI (red).

An alternative way of displaying the MW data is shown in Figure 5, which shows the MW change for individual compounds in going from the lead to drug plotted agains the MW of the final drug. It can clearly be seen that in the majority (78%) of the cases studied here, there is an increase in MW in the lead to drug process. As noted above, for the whole set there is an average MW increase of 38 for this data set. When only those molecules that increase in MW are considered, then the average increase is 63. MW is just one of several descriptors that can be correlated with the lead to drug process in this way. However it is not simply a matter of increase in molecular weight that relates to complexity as can be seen by the changes in other properties.

How then does the profile of leads and drugs developed above compare with that of the type of molecules that we tend to make in libraries. Figure 6 shows a series of ADEPT profiles for a selection of bead based libraries synthesized at Glaxo Wellcome in recent years. It can be seen that these profiles are typically significantly right shifted compared to the WDI (the red line with tick marks in Figure 6). This substantiates the hypothesis discussed above, that libraries are made up of molecules which are more complex compared to drugs and even more complex compared to leads.

Although our design criteria for libraries now incorporate concepts of drug likeness, it is still difficult to keep the properties to the drug-like profiles required.[5] The realization that we should be aiming our designs at leadlike profiles only makes this harder. Why is this so? It seems to be mainly as a result of the types of chemistry we use and the numbers of molecules that we believe are required in order to win at the "numbers game" that is inherent in the strategy of high throughput screening. We have seen that as molecular

Table 2. Average Property Values for the Sneader Lead Set, Average Change on Going to Sneader Drug Set, and Percentage Change[a]

av # arom	Δarom	%	av ClogP	ΔClogP	%	av CMR	ΔCMR	%
1.3	0.2[b]	15	1.9	0.5[b]	26	7.6	1.0[b]	14.5
av # HBA	**ΔHBA**	**%**	**av # HBD**	**ΔHBD**	**%**	**av # heavy**	**Δheavy**	**%**
2.2	0.3[b]	14	0.85	−0.05[c]	(4)	19.	3.0[b]	16
av MW	**ΔMW**	**%**	**av MV**	**ΔMV**	**%**	**av # rot B**	**Δrot B**	**%**
272	42.0[b]	15	289	38.0[b]	13	3.5	0.9[b]	23
av ABE	**ΔABE**	**%**	**av # bits**	**Δbits**	**%**			
8.3	2.6[b]	31	169	29[b]	17			

[a] Abbreviations: av # arom = average number of aromatic rings as defined by the Daylight software (v451); av ClogP = average calculated logP as defined by Daylight software (v451); av CMR = average calculated molar refractivity as defined by Daylgiht software (v451); Av Rot B = average number of rotatable bonds; av # HBA, av # HBD = average number of H-bond acceptors and doners as defined by in-house collection of SMARTS descriptors; av # heavy = average number of heavy atoms; av MV = average molecular volume (Schroedinger); av MW = average molecular weight; av ABE = average Andrew's binding energy; av bits = average number of bits set on in standard Daylight 1024 bitstring). [b] Paired sample *t*-test: highly significant (at < 0.1%). [c] Paired sample *t*-test: not significant.

complexity increases, there is an increasingly small chance of an individual molecule being a hit in an assay. But only complex molecules allow us to make the numbers we want (e.g. through several points of diversity). This is a classic catch-22 situation whereby we need more molecules to increase the odds, but these large numbers of molecules, as made via libraries, tend to be more complex. This actually decreases the chance of any individual compound being a hit in the first place. While it is difficult to prove, our experiences in lead discovery using libraries would suggest that the rate of decrease in the chance of finding a hit is much larger than the increase in the number of molecules that result from making more complex molecules.

HIGH CONCENTRATION SCREENING AND MULBITS

One way to break this cycle is to work with less complex molecules. However, less complex molecules are likely to express lower binding affinities, which consequently requires the use of robust assays that allow for higher ligand concentration screening in order to observe weaker binding. It is also useful if an insight can be gained into the specificity or multiplicity of binding by the use of some appropriate biophysical technique. An extreme example of this is illustrated by an approach that we developed some years ago termed MULBITS. As an example of the use of MULBITS, in our thrombin inhibitor program we developed an assay that could detect small MW entities binding exclusively in the S1 pocket. It has previously been shown by X-ray analysis that proflavin binds exclusively in this pocket and that displacement of proflavin by ligands (as detected by absorbance of proflavin) provides the basis for a simple assay that is specific to this region of the protein.[26] In this way we were able to rank a range of likely and novel P1 substituents prior to their synthetic incorporation into larger molecules. An example of one such MULBIT discovered in this way was 2-aminoimidazole, and the final molecule that this was incorporated into is shown in Figure 7. In this case the complexity of the synthesis of the translactam system made it highly desirable to prioritize possible P1 substituents in advance of committing chemistry resource.[27]

The SAR by NMR work of Fesik has obvious similarities to this approach except that this uses NMR as the method of detection.[29] More recently Ellman and colleagues have published a related methodology that utilizes the capped monomers of potential combinatorial libraries as the MULBITS for high concentration screening (typically at concentrations of 500 μM).[30] Another example, concerning inhibitors of DNA-gyrase, has recently appeared in the literature.[31]

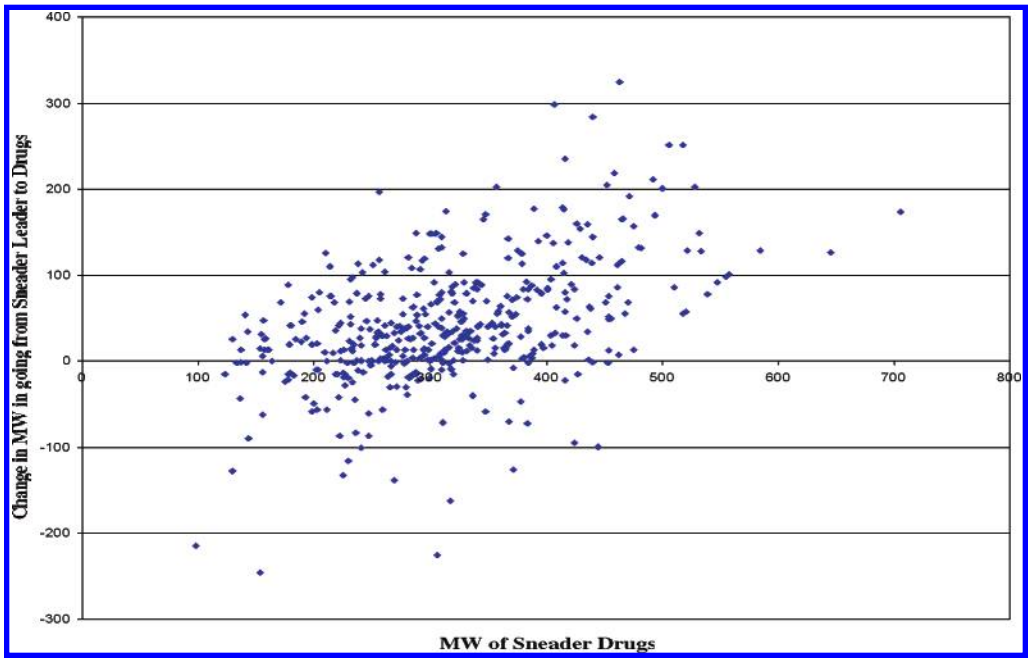

Figure 5. Change in MW in going from Sneader lead to drug as a function of MW of final drug.

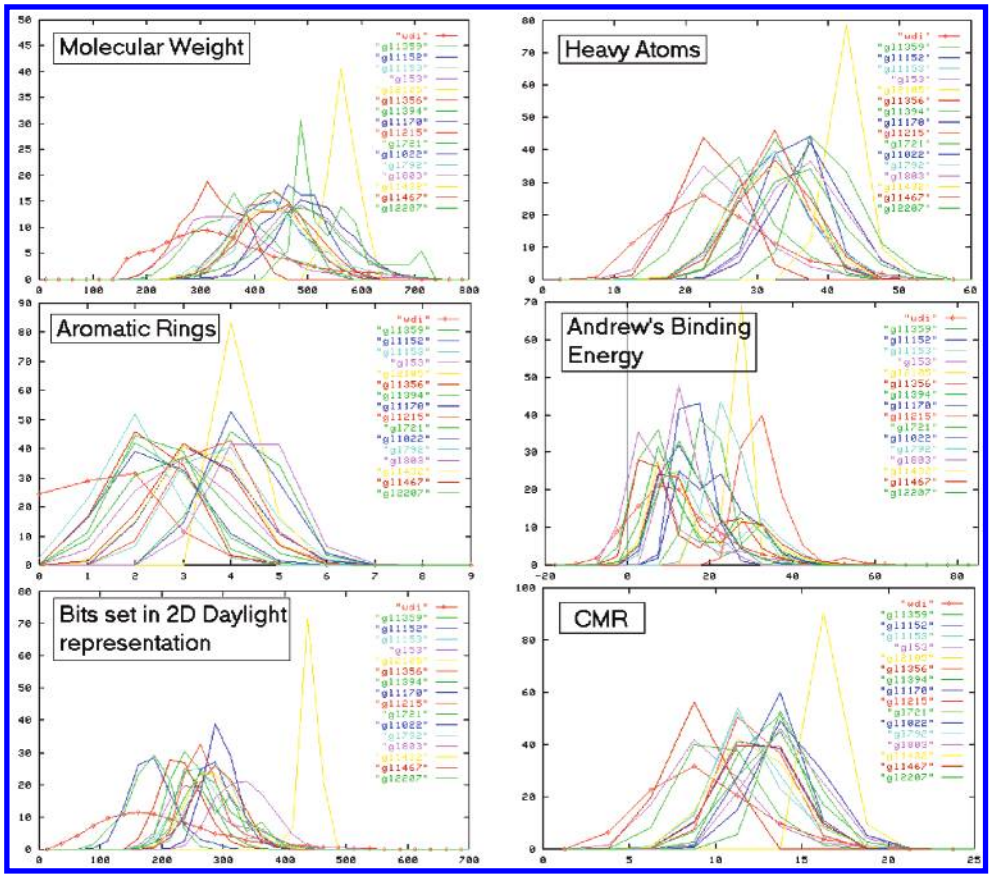

Figure 6. ADEPT profiles for WDI (red with * on line) and 16 libraries.

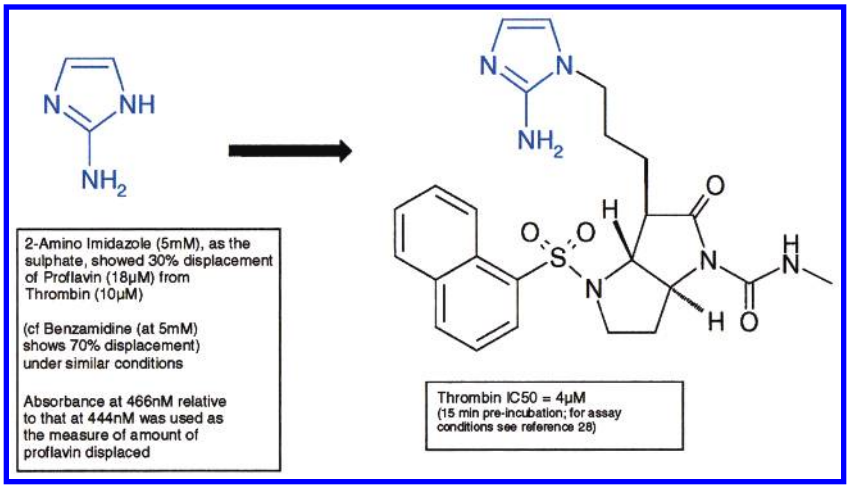

Figure 7. The discovery of 2-aminoimidazole as an S1 MULBIT for thrombin and its incorporation into a more complex ligand.

It uses a combination of many of the above concepts. All of these methods are designed to break the combinatorial explosion problem by trying to find initial leads (or monomers for libraries' work) by reducing the search space (i.e. by screening initially in the equivalent of monomer space). They also fulfill the requirements discussed above, in that the initial leads are of reduced complexity. These MULBITS type methods rely on finding "usable" activity (200 μM to 10 mM) in small molecules (typically 100−250 Daltons). From this initial lead it is then possible to optimize activity and other properties by increasing the complexity. This form of prescreening in monomer space is a highly efficient and cost-effective way of assessing the diversity of a much larger number of molecules. Such methods are worth considering as complementary to standard HTS where molecules that are of greater complexity are routinely used. This is because, for the reasons discussed, the standard HTS approach is flawed in some situations.

CONCLUSIONS

Our exploration of simple models of the probability of ligands interacting effectively with receptors provides a novel framework within which to discuss the nature of drug leads. When coupled with results from profiling a series of drug discovery histories and our own and other workers' experiences of using molecules with reduced complexity for screening, this approach suggests a highly complementary approach to that of HTS for discovering leads. It also

challenges the current design philosophies around combinatorial libraries and the propensity for such libraries to yield overly complex molecules. Albert Einstein is credited with the saying "everything should be made as simple as possible, but not simpler". This makes an appropriate rule for the design and selection of molecules as potential leads to be found by high throughput screening.

ACKNOWLEDGMENT

We thank John Bradshaw and Colin Grey for converting the Sneader dataset into a Daylight database, Vipal Patel for synthesis of the Thrombin ligand, and Sue Bethell and Charlie Nichols for the development of the Proflavin Thrombin assay. We are most grateful to John Bradshaw, Giampa Bravi, Andy Brewster, Robin Carr, Miles Congreve, Darren Green, Brian Evans, Albert Jaxa-Chamiec, Duncan Judd, Xiao Lewell, Mika Lindvall, Steve McKeown, Adrian Pipe, Nigel Ramsden, Derek Reynolds, Barry Ross, Nigel Watson, Steve Watson, and Malcolm Weir for helpful discussions that have contributed to this work

REFERENCES AND NOTES

(1) Hird, N. Isn't combinatorial chemistry just chemistry? *Drug Discovery Today* **2000**, *5(8)*, 307−308.
(2) Leach, A. R.; Hann, M. M. The *in silico* world of virtual libraries. *Drug Discovery Today* **2000**, *5(8)*, 326−336.
(3) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23(1−3)*, 3−25.
(4) Sadowski, J.; Kubinyi, H.. A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* **1998**, *41(18)*, 3325−3329
(5) Gillet, V. J.; Willett, P.; Bradshaw, J.; Green, D. V. S. Selecting Combinatorial Libraries to Optimize Diversity and Physical Properties. *J. Chem. Inf. Comput. Sci.* **1999**, *39(1)*, 169−177.
(6) Teague, S. J.; Davis, A. M.; Leeson, P. D.; Oprea, T. The design of leadlike combinatorial libraries. *Angew. Chem., Int. Ed. Engl.* **1999**, *38(24)*, 3743−3748.
(7) Leach, A. R. *Molecular Modeling: Principles and Applications*; Longmans: 1996; pp 480−541.
(8) Andrews, P. R.; Craik, D. J.; Martin, J. L. Functional group contributions to drug-receptor interactions. *J. Med. Chem*. **1984**, *27-(12)*, 1648−57.
(9) Williams, D. H.; Westwell, M. S. Aspects of weak interactions. *Chem. Soc. Rev.* **1998**, *27(1)*, 57−64.
(10) Bissantz, C.; Folkers, G.; Rognan, D. Protein-Based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations. *J. Med. Chem.* **2000**, *43(25)*, 4759−4767.
(11) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *J. Am. Chem. Soc.* **1997**, *119(43)*, 10509−10524. (b) Venkatarangan, P.; Hopfinger, A. J. Prediction of Ligand−Receptor Binding Free Energy by 4D-QSAR Analysis: Application to a Set of Glucose Analogue Inhibitors of Glycogen Phosphorylase. *J. Chem. Inf. Comput. Sci*. **1999**, *39(6)*, 1141−1150.
(12) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-Point Pharmacophore Method for Molecular Similarity and Diversity Applications: Overview of the Method and Applications, Including a Novel Approach to the Design of Combinatorial Libraries Containing Privileged Substructures. *J. Med. Chem*. **1999**, *42(17)*, 3251−3264.
(13) Rejto, P. A.; Verkhivker, G. M. Unraveling principles of lead discovery: From unfrustrated energy landscapes to novel molecular anchors. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93(17)*, 8945−8950.
(14) A tanh function was used to create this shape of curve.
(15) Cochran, A. G. Antagonists of protein−protein interactions. *Chem. Biol*. **2000**, *7(4)*, R85−R94.
(16) Sneader, W. *Drug Prototypes and their Exploitation*; John Wiley and Sons Ltd.: 1996.
(17) Daylight theory manual and related documentation available at http://www.daylight.com/dayhtml/doc/theory/theory.toc.html.
(18) Analysis showed these natural product derived compounds to have a much wider distribution of properties and changes on going from lead to drug, reflecting that natural products provide a different (but effective way) in which drugs have been developed. The fact that natural products have been arrived at over many millennia of evolution suggest that considerable optimisation of interactions with related biotargets has already be embedded in the leads that we have then been found via screening.
(19) Leach, A. R.; Bradshaw, J.; Green, D. V. S.; Hann, M. M.; Delany, J. J., III. Implementation of a System for Reagent Selection and Library Enumeration, Profiling, and Design. *J. Chem. Inf. Comput. Sci*. **1999**, *39(6)*, 1161−1172.
(20) World Drug Index published by Derwent available from Daylight Inc.
(21) Calculated logP as calculated by PC-models module from Daylight Inc; http://www.daylight.com/.
(22) Number of aromatic rings calculated using Daylight toolkit.
(23) Andrew's binding energy calculated via in-house routines based on ref 8.
(24) Number of bits set on in the standard 1024 Daylight fingerprint. This reflects the number of different atoms and bond connection within a molecule and is therefore a useful measure of the internal complexity of a molecule.
(25) Calculated Molecular Refractivity (CMR) as calculated by PC-models module from Daylight Inc; http://www.daylight.com/.
(26) Conti, E.; Rivetti, C.; Wonacott, A and Brick, P. X-ray and spectrophotomeric studies of the binding of proflavin to the S1 specificity pocket of human thrombin. *FEBS Lett.* **1998**, *425*, 229−233.
(27) Coote, S. J.; Dowle, M. D.; Finch, H.; Hann, M. M.; Kelly, H. A.; MacDonald, S. J. F.; Pegg, N. A.; Ramsden, N. G.; Watson, N. S.. Furopyrrolidine derivatives and their use as serine protease inhibitors. *PCT Int. Appl*. **1999**, WO 9912936.
(28) Pass, M.; Abu-Rabie, S.; Baxter, A.; Conroy, R.; Coote, S. J.; Craven, A. P.; Finch, H.; Hindley, S.; Kelly, H. A.; Lowdon, A. W..; McDonald, E.; Mitchell, W. L.; Pegg, N. A.; Procopiou, P. A.; Ramsden, N. G.; Thomas, R.; Walker, D. A.; Watson, N. S.; Jhoti, H.; Mooney, C. J.; Tang, C.; Thomas, P. J.; Parry, S.; Patel, C. Thrombin inhibitors based on [5, 5] trans-fused Indane lactams. *Bioorg. Med. Chem. Lett.* **1999**, *9(12)*, 1657−1662.
(29) Hajduk, P. J.; Meadows, R. P.; Fesik, S. W.. Discovering high-affinity ligands for proteins. *Science* **1997**, *278(5337)*, 497−499.
(30) Maly, D. J.; Choong, I. C.; Ellman, J. A.. Combinatorial target-guided ligand assembly: identification of potent subtype-selective c-Src inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97(6)*, 2419−2424.
(31) Boehm, H.-J.; Boehringer, M.; Bur, D.; Gmuender, H.: Huber, W.; Klaus, W.; Kostrewa, D.; Kuehne, H.; Luebbers, T.; Meunier-Keller N.; Mueller, F. Novel Inhibitors of DNA Gyrase: 3D Structure Based Biased Needle Screening, Hit Validation by Biophysical Methods, and 3D Guided Optimization. A Promising Alternative to Random Screening. *J. Med. Chem.* **2000**, *43(14)*, 2664−2674.

CI000403I