J Comput Aided Mol Des (2013) 27:783–792
DOI 10.1007/s10822-013-9683-1

# The importance of molecular complexity in the design of screening libraries

Shahul H. Nilar · Ngai Ling Ma · Thomas H. Keller

Received: 26 May 2013 / Accepted: 7 October 2013 / Published online: 13 October 2013

**Abstract** The one-dimensional model of Hann et al. (J Chem Inf Comput Sci 41(3):856–864) has been extended to include reverse binding and wrap-around interaction modes between the protein and ligand to explore the complete combinatorial matrix of molecular recognition. The cumulative distribution function of the Maxwell–Boltzmann distribution has been used to calculate the probability of measuring the sensitivity of the interactions as the asymptotic limits of the distribution better describe the behavior of the interactions under experimental conditions. Based on our model, we hypothesized that molecules of lower complexity are preferred for target based screening campaigns, while augmenting such a library with moieties of moderate complexities maybe better suited for phenotypic screens. The validity of the hypothesis has been assessed via the analysis of the hit rate profiles for four ChemBL datasets for enzymatic and phenotypic screens.

**Keywords** Molecular complexity · Selectivity · Sensitivity · Intermolecular interaction

## Introduction

The identification of lead molecules is one of the most important steps in the drug discovery process. Often the quality of the chemical starting point is a major determinant for the success or failure of a lead optimization program. A number of technologies have been developed in the past few years to facilitate this process, all of them relying on the measurement of the physical interaction between proteins and small molecular weight compounds. The dominant technology is still high throughput screening (HTS) but fragment based screening is fast becoming a viable alternative using X-ray and nuclear magnetic resonance (NMR) techniques [1, 2].

There is no question that target based approaches are the state of the art in drug discovery. However, for various reasons, a number of infectious disease indications are still largely relying on cellular screening as the major lead finding approach [3, 4]. This is especially true for bacterial diseases where high throughput screening of bacterial proteins has repeatedly failed to yield tractable lead molecules [5]. Since target based approaches have dominated industrial drug discovery, we have considerable information to guide us in choosing the best libraries for lead finding [6–8]. The situation is completely different for phenotypic screening. While natural products have provided a number of drugs from such approaches [9], it is not clear whether compounds from natural sources are particularly well suited for cellular screening or whether libraries of synthetic compounds (which are much more readily available) would be useful in this context [10].

A cellular system can be viewed as an ensemble of enzymes. Chemical intuition suggests that compounds with a Lipinski [11] or lead-like profile [12] will interact with a number of such enzymes due to viable interactions between the protein and the small molecule. As the potency of a cellular hit is a composite of all the productive interactions of a compound inside a cell, promiscuous compounds [13, 14] are likely to produce structure–activity relationships that

S. H. Nilar (✉) · N. L. Ma · T. H. Keller
Novartis Institute for Tropical Diseases, 10 Biopolis Road,
#05-01 Chromos, Singapore 138670, Singapore
e-mail: shahul.nilar@novartis.com

*Present Address:*
T. H. Keller
Experimental Therapeutic Center, Agency for Science,
Technology and Research, 31 Biopolis Way, Singapore 138669,
Singapore

are difficult to interpret and therefore pose challenges in lead optimization campaigns (i.e. such compounds are unattractive chemical starting points). This thought process leads to the question whether it is possible to bias screening libraries for cellular screening towards compounds that are more likely to interact with a limited number of protein targets inside a cell.

While compounds isolated from natural sources have a number of unique features [15], the most interesting property of these molecules in the context of lead finding is structural complexity. Several studies have examined whether structural complexity plays an important role in target based lead finding [16, 17]. However it would be interesting to examine the topic in a wider context by asking: is there any relation between complexity (molecular or structural) and library design for target-based/phenotypic/fragment based screening?

While chemists share the general view that molecules are complex, providing a precise definition of molecular complexity is subjective. In this paper, the idea of molecular/structural complexity is closely associated with the number of features on the ligand and the region(s) of interest in the target protein(s) that are conducive to interactions that result in measurable binding. To better understand these issues, an analysis of molecular complexity within the framework of molecular recognition within a simple one-dimensional model is presented. This approach is different from the graph theoretical examinations of complexity [18, 19]—in the approach presented here, the fundamental intermolecular interactions of ***one*** compound with ***one/many*** proteins are used to mimic enzymatic/phenotypic screenings, respectively.

The analysis presented here is based on the seminal work of Hann et al. [20] in which one dimensional protein/ligand pairs are represented as strings of interaction points with each point assigned a+ or −sign. Our work extends the Hann model and seeks to explain the probability of obtaining a hit/lead in terms of the selectivity and complexity of the chemical entity and the sensitivity of the measuring process through the inclusion of an exhaustive description of the protein and ligand interactions. Our model includes reverse modes, wrap around modes, and allows for mismatch between the protein and the ligand, which were excluded in the analysis presented in Ref. [20]. Using this analysis we have attempted to better understand the role of complexity for library design.

## Methods

Intermolecular recognition arises from the complementary matching of features on the molecules involved. Such features typically include hydrogen bonding interactions, centers of hydrophobic interactions, columbic interactions, polarizabilities and derivatives of these terms, each with different contributing strengths towards recognition. In the Hann model [20], the protein and ligand are each represented as a one-dimensional array of interaction sites. Although this approach oversimplifies the physics behind the intermolecular interactions involved in recognition, it has been shown [6, 20, 21] that the model provides a useful conceptual framework to understand issues concerned with library design for lead finding approaches.

Using terminology similar to that in Ref [20], the number of interaction sites for protein ***P*** and ligand ***L***, is denoted as *p* and *l*, respectively, in the work presented here. The complexity of each moiety is thus defined as the constituent number of such interaction points. The one-dimensional depiction of the protein and ligand as an array of interaction sites, with each site depicted with a+ or − sign is shown in Fig. 1 for a protein and ligand of complexities 8 and 3. A *match between the protein and the ligand* is defined as *the overlap of interaction sites of opposite polarities i.e.* a match is said to occur when sites with the opposite signs are juxtaposed (as depicted in Fig. 1). The Hann model allows only for ***complete*** matches between the protein and the ligand, in which ***all*** the interaction points of the ligand overlaps with complementary points of opposite signs in the protein. While such a treatment simplifies the analysis, it ignores the possibilities of ***partial*** matches in which only sections of the ligand are matched with the protein leading to an underestimation of the overall interaction. The inclusion of partial matches is especially important for ligands with higher complexity (for example natural products), since the probability of having a ***complete*** match to a protein will be very low. Furthermore, reverse binding and wrap-around modes (defined in the section below) for the ligand are discounted in the Hann model, and the inclusion of which is necessary for a more complete representation of the interaction space for molecular recognition:

Binding interactions between the protein and the ligand are defined as follows:

1. An interaction is defined to occur when a + sign on the ligand interacts with a − sign on the protein or the other way round (these matches are designated with arrows in Fig. 1).
2. A binding mode is one in which the described representation of the protein and the ligand comprise a combination of signs that are complementary. An *m*-site binding mode occurs when there are *m* allowed interactions between the protein and ligand (Fig. 1a–e)
3. Reverse binding mode, in which the ligand binds in the different orientation ("− ++" in Fig. 1d, compared to "++ −" in Fig. 1a, b, c).

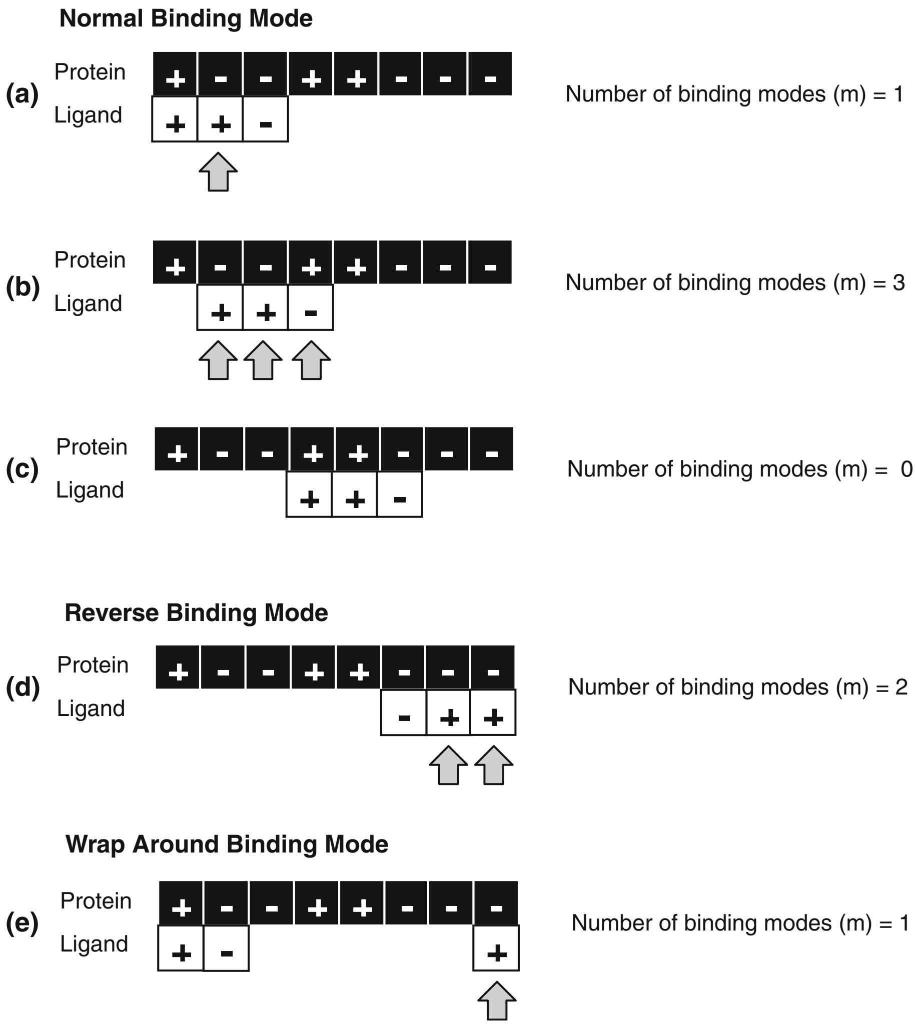

**Fig. 1** Definition of the binding interactions for a protein of complexity 8 and a ligand of complexity 3. The *arrows* denote a matched interaction and the mismatched sites on the ligand have the same polarity on the protein

4. Wrap around binding interactions, where the ligand interacts with “non-consecutive” parts of the protein (in 1-dimension), is shown in Fig. 1e. Such interactions are plausible given the 3-dimensional nature of the protein.

A protein $\boldsymbol{P}$ can be written in the set notation as $\boldsymbol{P} = \{(+ \text{ or } -)_j \mid j = 1, \ldots, p\}$. There are $2^p$ such combinatorial representations for each protein $\boldsymbol{P}$ and within this notation, and the interaction site of interest (active or binding site) can be any one of these possibilities. Similarly, a ligand $\boldsymbol{L}$ can have $\boldsymbol{l}$ interaction sites resulting in $2^{\ l}$ such combinations. This representation for the protein and the ligand is termed the “polar representation” in this paper for convenient reference. A successful protein–ligand interaction is defined as the combinatorial overlap between a site on the protein within the polar representation with a site of opposite polarity on the ligand (Fig. 1).

The mathematical basis of this problem compels the use of combinatorial algebra techniques that have been utilized in the derivation of Eqs. (3)–(5) below.

The Hamiltonian $\hat{H}$ for the protein–ligand system is given by.

$$\hat{H} = \hat{H}_P + \hat{H}_L + \hat{H}_{P-L} \tag{1}$$

where the subscripted components $P$, $L$ and $P{-}L$ refer, respectively, to the Hamiltonian operators for the protein, ligand and the interaction of the protein with the ligand. The total energy $\boldsymbol{E}$ of the composite system is given by:

$$E = E_P(f(p_i)) + E_L(f(l_i)) + E_{P-L}(p_1 p_2 \cdots p_p \otimes l_1 l_2 \cdots l_l) \tag{2}$$

where the subscripts $P$, $L$ and $P{-}L$ have the same meaning as explained above. The terms $E_P(f(p_i))$ and $E_L(f(l_i))$ denote the energies of the non-interacting protein and ligands, respectively, which are functions of the constituent interaction points. The energy term of interest in this work is the third term in Eq. (2), $E_{P-L}(p_1 p_2 \ldots p_p \otimes l_1 l_2 \ldots l_l)$, which represents the protein–ligand interaction energy. As each string of interaction points is represented by a combination of $+$ or $-$ signs, the multiplier $\otimes$ assigns a value of unity for the overlap of each pair of oppositely signed interaction points (Fig. 1). The number of binding modes, ***m***, is the sum of the number of such matches between the interacting moieties.

The number of ways ***w***, in which a protein and ligand can have ***m***-matches within this representation is given by:

$$w = \overset{p}{\underset{l}{C}}\,\overset{l}{\underset{m}{C}} \tag{3}$$

The first combinatorial term of Eq. (3), $\overset{p}{\underset{l}{C}}$, is the number of ways a protein and a ligand can combine. Given the particular choice of the protein–ligand combination within the polar representation, the second term $(\overset{l}{\underset{m}{C}})$ gives the number of ways in which the ligand can interact with the protein with ***m*** points. Two cases are highlighted below. In the first, the number of interaction sites on the protein and the ligand are the same i.e. $\boldsymbol{p = l}$. In the second case, there exists an inequality in the number of interaction sites, usually with $\boldsymbol{p > l}$**,** which corresponds to a more physically plausible scenario. There are examples reported for which the reverse inequality holds true, (i.e. $\boldsymbol{p < l}$) as in the case of the Zanamivir binding pocket [21] and the binding of the 'starfish' like carbohydrate ligand to Shiga-like toxins [22].

There are $2^{\,l}$ polar representations for the ligand. Thus, the probability of an *m*-site binding between the protein and the ligand is given by:

$$\text{Selectivity} = Pr\,[m - \text{site binding}] = \frac{\overset{p}{\underset{l}{C}}\,\overset{l}{\underset{m}{C}}\,2^{l}}{2^{p+l}} = \frac{\overset{p}{\underset{l}{C}}\,\overset{l}{\underset{m}{C}}}{2^{p}} \tag{4}$$

Within the one-dimensional model employed here, the probability of the interaction between a protein and a ligand is given by Eq. (4), and defined as the ***selectivity*** between these two species.

Experimentally, the process of measuring molecular recognition phenomena involves an assay process. The result from such a measurement is the product of the ***selectivity*** of the interactions between the molecular entities in the assay and the ***sensitivity*** of the measuring process itself. In keeping with Hann's work [20], the probability *Pr* of measuring a useful binding event is defined as:

$$\begin{aligned} &Pr[\text{measuring a useful binding event}] \\ &\quad = Pr[\text{protein - ligand matches}] \times Pr[\text{measuring binding}] \end{aligned}$$

where the first term on the right hand side is a measure of the selectivity of the intermolecular recognition process and the second term being the sensitivity of the measuring process.

The sensitivity of the measuring process is interpreted as the probability of obtaining a measurement from the experiment, such as an assay process. This property should include the various intermolecular binding interactions, such as the partially matching pairs, reverse binding and the wrap-around modes, in the molecular recognition process observed during the time span of the experiment. The cumulative distribution function (CDF) of the Maxwell–Boltzmann distribution has been used here to represent this probability of the measuring process and is shown in Eq. (5). In the Hann paper [20], the sensitivity was represented by the hyperbolic tangent function of the ligand complexity. Although the shape of the function is similar to that of the CDF used here, the asymptotic behavior differs, with the hyperbolic tangent approaches unity for ligands with complexity ≥8 (Ref. [20], Fig. 3), thereby empirically biasing compounds to this complexity range. In using the CDF of the Maxwell–Boltzmann distribution in this communication, the asymptotic behavior is now associated with compounds of much higher complexity and the sensitivity of the experimental process is dependent on ***both*** the ligand complexity and the extent of interaction between the protein and the ligand. Furthermore, the statistical nature of the molecular recognition process is established by linking the sensitivity of the experiment with the Boltzmann distribution.

The functional form of the CDF used in this paper is:

$$Pr(\text{measuring binding}) = erf\left(\frac{m}{\sqrt{2}l}\right) - \sqrt{\frac{2}{\pi}}\frac{m}{l}\exp\left(\frac{-m^{2}}{2l^{2}}\right) \tag{5}$$

where ***l*** and ***m*** are the ligand complexity and the number of binding mode interactions between the protein and the ligand, respectively. The error function *erf* (x) is defined in the usual manner as:

$$erf(\mathrm{x}) = \frac{2}{\pi}\int\limits_{0}^{x} \exp\left(-t^{2}\right)dt$$

In the non-binding event ($\boldsymbol{m} = 0$), Eq. (5) returns a value of zero for the probability of measuring binding which reflects the experimental observation. The numerical behavior of the CDF ranges from 0 at small values of ***m*** and ***l*** and approaches unity at high values for the ligand complexity in a sigmoidal manner. Thus, the use of the cumulative distribution function as a measure of the sensitivity indicates a finite probability of measuring a molecular recognition process (i.e. a binding event) if the period of observation is infinite, or, over a very long time span.

Combining Eqs. (4) and (5), the probability of measuring a useful binding event of order *m* (i.e. *m* points of interaction between the protein and the ligand) is given by:

$$Pr(\text{measuring a useful binding event}) = erf\left(\frac{m}{\sqrt{2l}}\right) - \sqrt{\frac{2}{\pi}}\frac{m}{l}\exp\left(\frac{-m^2}{2l^2}\right)\cdot\frac{\overset{p}{\underset{l}{C}}\overset{l}{\underset{m}{C}}}{2^p} \quad (6)$$

It is this equation that has been used to evaluate the probability profiles for protein–ligand interactions discussed in the next section.

As highlighted in the introduction, a precise definition of molecular complexity is subjective. In Hann et al. [20] the Sneader [23] set of drugs were compared with those entities listed in the World Drug Index (WDI) in terms of the molecular weight, clogP and other physicochemical properties such as the number of aromatic rings and heavy atoms. Schuffenhauer et al. [24] analyzed the Novartis High-Throughput Screening (HTS) summary data and results from screening compound libraries using nuclear magnetic resonance (NMR) techniques, using the number of SIMILOG [25] keys present in the molecule. In this paper, molecular complexity was defined as the total number of Klekota-Roth molecular keys [26] present in each molecule. By mining the results generated from 24 screening campaigns [26] on a 16,203 diverse compound set together with the data from screening 37,330 compounds for growth inhibition against 70 cancer cell lines (NCI), 4,860 unique substructures were generated. These substructures were obtained by fragmenting the compounds, collecting the most abundant and the most discriminating moieties that are associated with the biological activity in the datasets.

Four ChemBL [27] datasets were analyzed in this work. The datasets used as examples of enzymatic screens were for kinases and G-protein coupled receptors (GPCR). Only those data points that had IC50 data were retained for each dataset. Duplicate compounds with identical assay were removed, resulting in 70,935 data for the kinase set and 67,166 compounds for the GPCR library. For the analysis of data from phenotypic screens, the GlaxoSmithkline Tres Cantos Antimalaria dataset [28] which contains does response inhibition data for 13,404 compounds against the Malaria *Plasmodium Falciparum* and the World Health Organization WHO–TDR data for the screening of 740 compounds against Human African Trypanosomiasis (HAT), Leishmaniasis and Chagas diseases were used (CHEMBL2093137) [29]. The compounds in each data set were binned into categories based on the logarithm of the inhibitory dose response in increments of unity. In this work, hit rate was defined as the number of molecules that belong in each bin of activity divided by the total number of molecules in the dataset analyzed. The results of the analysis are presented in Figs. 5, 6 below.

## Results and discussion

For a set of proteins and ligands of complexity varying from 2 to 8, the probability of measuring binding events were calculated using Eq. (6) above. It is important to emphasize that for each value for the complexity of the protein (ligand), the number of polar representations is $2^p$ ($2^{\,l}$), respectively. Thus, each profile for a given value of complexity reflects the behavior of an ensemble of interacting moieties.

In Fig. 2, the binding profiles for a protein of complexity 8 (i.e. a protein with 8 possible sites of interaction) are shown as a function of ligand complexity. Each profile has been calculated for $m$ binding modes, where $m = 0$–8, where the zeroth order binding mode is the probability of measuring a non-binding event; $m = 1$–7 corresponds to partial matches; and $m = 8$ corresponds to complete perfect matches. Visual analysis of the profiles in Fig. 2 shows that the probability of measuring a partial protein–ligand match ($m < 8$) is much higher than that of observing a perfect binding (perfect match) event ($m = 8$). In other words, a perfect match in which all the +/− interactions between the protein and the ligand are satisfied is rare. Inclusion of partial interactions between the ligand and the receptor together with reverse and wrap-around modes shows that the maximum probabilities are observed when the number of recognition points between the moieties is 3 or 4 for ligands having complexities of 4 or 5. Experience in pharmacophore based applications has shown that 4–5 features results in an optimum pharmacophore that describes protein–ligand interactions [30–32]. However, extrapolation of the results described in this paper to such

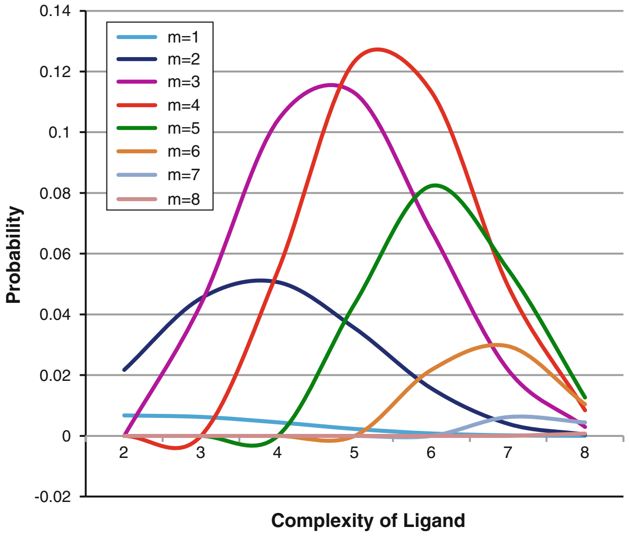

**Fig. 2** Binding profiles for a protein of complexity 8 as a function of ligand complexity

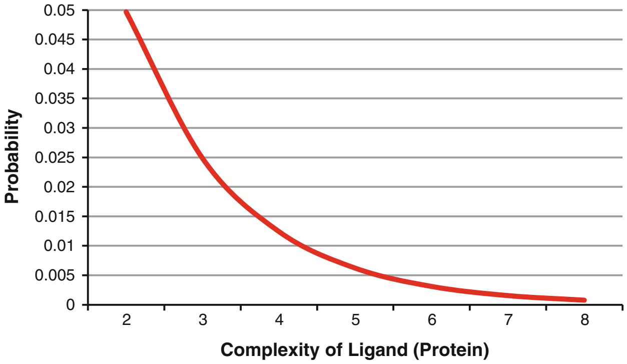

**Fig. 3** Selectivity and the maximum probability of measuring the binding of perfect protein–ligand matches as a function of ligand (protein) complexity

conclusions are viewed with caution as the model employed here is conceptual and does not include all of the physics that define molecular recognition processes. As the level of complexity increases, the number of possible polar representations increases exponentially, thereby increasing the number of non-interacting mismatches, which reduces the probability of binding. It is also interesting to note from Fig. 2 that ligands with complexity of 4 or less are predicted to have a measurable probability in an assaying campaign indicating the importance of compounds of lower complexity in a screening library.

While similar qualitative conclusions were drawn if only complete interactions are included [20], the case of ligands with higher complexity measures warrant special attention. Natural products often exhibit high molecular/structural complexity which usually arises from their size, chirality and the availability of many interaction sites for hydrogen bonding. The probability of measuring the binding of these compounds is usually attributed to a section of the molecule interacting with the target site [33], which is supported by our model. Figure 3 shows the maximum probability values of measuring protein–ligand interactions for a perfectly matching molecular recognition process. The exponential decay in the probability of finding perfect matches as the ligand complexity increases shows that such a measurement for a highly complex ligand such as a natural product will be rarely observed. This in turn suggests that the measurable binding of natural products in a screening experiment, is mainly due to the partial modes of binding in the intermolecular recognition processes. It is the inclusion of such partial modes in this work, via a combinatorial approach, that allows the model to rationalize such observations.

The current model, in agreement with Hann [20], suggests that molecules of lower complexity have a higher selective tendency to bind at receptor sites. This is in keeping with the paradigm shift towards lead libraries of

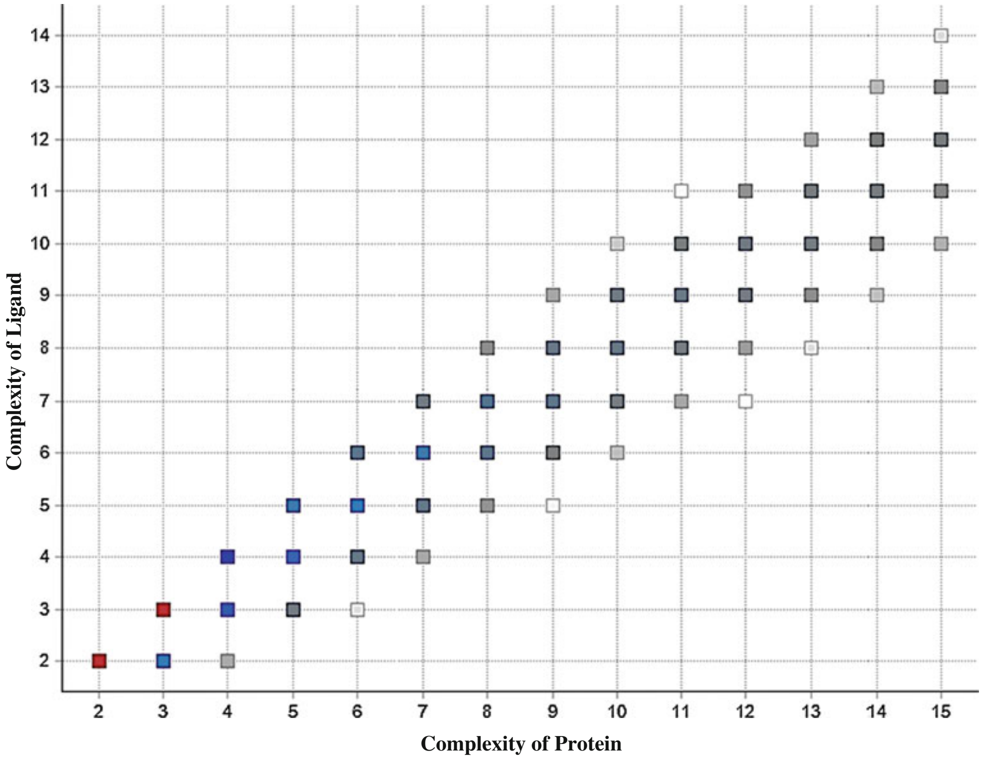

**Fig. 4** Probability distributions for proteins with complexities, ***p***, (*x-axis*), ligand complexities, ***l***, (*y-axis*) each ranging from 2 to 15. Higher values for the probabilities are colored in *red*, with the medium and low probabilities represented in *blue* and *grey* respectively

smaller compounds or fragments [33–38]. From Eq. (6) above, the sensitivity of measuring the binding of molecules to receptors is dependent on the binding order ***m*** between the interacting moieties. Thus, fragments and molecules of lower complexity will have a lower sensitivity as predicted by Eq. (6) as is observed under experimental conditions.

Even in today's environment where rational avenues towards library design have been explored, most lead finding campaigns embrace a strategy that calls for the screening of as many compounds as possible (often more than a million compounds) [38, 39]. While such high throughput screening campaigns are quite expensive, the outcome in terms of tractable hits is often sobering, especially for non-traditional targets. Screening libraries are historical collections and usually contain compounds in the evolution of the drug discovery strategies within the corporate environment [40]. As most of these compounds are of high complexity optimized for specific targets, is it realistic to expect these them to show a measurable degree of selectivity to other unrelated targets? In other words, is it optimal to screen the ***same*** library of compounds for ***both*** enzymatic and cell based screens?

A cell-based screen can be approximated as a collection of proteins of different complexities in a cell. This can be expressed as:

$$\Pi = \sum_{i=1}^{k} w_i P_i \qquad (7)$$

where Π and P represent the entire cell based system and the individual proteins in the system respectively. The contribution of each protein $P_i$ is weighted by $w_i$ and the summation runs over all the k-proteins that comprise the system. Thus, in a phenotypic screen, the probability profile for a measurable binding event of a molecule (ligand) of given complexity against the various enzymes in the cell will be an overlap of probability contours similar to those shown in Fig. 2. Overlapping the probability profiles for protein and ligand complexities, each moiety having complexity in the range of 2–15, gives rise to Fig. 4. Within the one-dimensional model discussed in this paper, the figure represents a generalized view of screening results. Each point on the figure corresponds to the probabilities of molecular recognition for protein–ligand of given complexity. Viewing the probability distribution across a horizontal grid in this Figure corresponds to the screening of compounds of a known complexity against a spectrum of complex proteins. A composite view of all the horizontal grids thus corresponds to cell based screen results. Similarly, the vertical grid view of Fig. 4 illustrates the probability distribution of enzymatic or target based screen, i.e. for a protein of fixed complexity, the library of compounds screened comprise molecules of different complexities. In Fig. 4, the probabilities of molecular recognition for protein–ligand of given complexity are color-coded red (high), blue (medium), and grey (low), respectively. The higher probabilities (colored red and blue) of interaction are observed for ligands of low and medium complexity for proteins of varying degrees of interaction sites. Furthermore, as the protein complexity increases, ligands of medium complexity emerge as having recognizable probabilities (colored blue in Fig. 4). Thus, in a phenotypic screen, where proteins of different levels of complexities are present, including ligands of medium complexity may increases the chances of obtaining hits from a screening process.

In Fig. 5, the hit rates are plotted for the Kinase and GPCR datasets in terms of molecular complexity, defined as the total count of the presence of the Klekota-Roth keys in the molecule. As discussed in this section, based on the analysis of the model presented in this work, a higher hit rate is expected for compounds of lower complexity. Indeed, this is observed for both the GPCR and Kinase data sets analyzed here. In the analysis of the GPCR data,

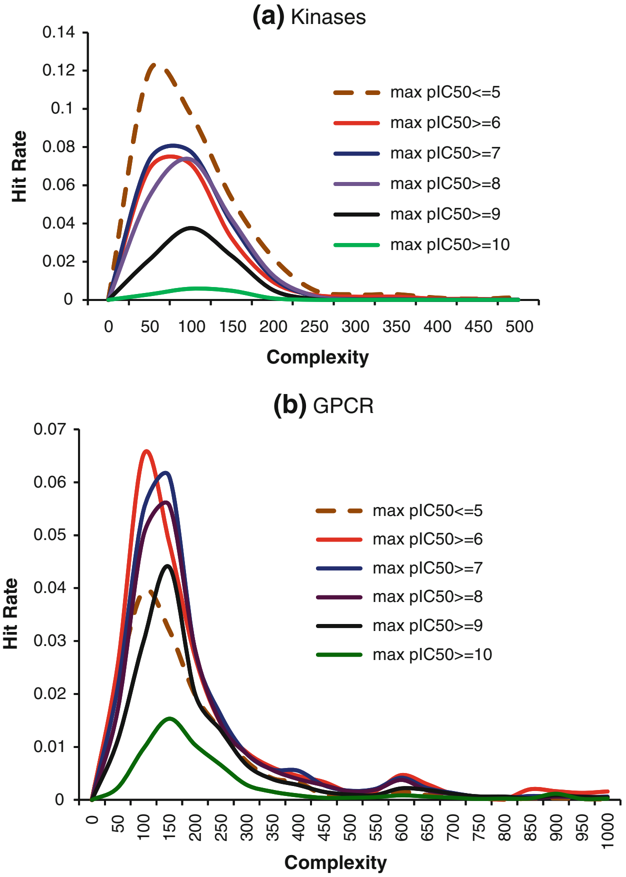

**Fig. 5** Hit rates for kinases (**a**) and GPCR (**b**) as a function of molecular complexity

molecules in which the number of Klekota-Roth keys are in the range [150–200] have a higher hit rate than those with higher complexities. For the Kinase inhibitors, a similar trend holds with the maximum value for the complexity centered around 100.

The inhibition data for Malaria Plasmodium Falciparum reported in the GSK-Tres Cantos lists only those compounds that are more potent than 1 µM. Consequently, obtaining a more complete profile of the hit rate dependence with molecular complexity in terms of the lesser active compounds is not possible with this dataset. However, the WHO–TDR data on Chagas, HAT and Leishmaniasis contains data on the compounds that are less than 1 µM inhibitory activity against the parasites. Comparing the hit rate of these two cellular systems, compounds with lower complexities have a higher hit rate than those with large values for this measure (Fig. 6). The contours in Fig. 6b for the WHO–TDR data clearly show that compounds with complexities centered around 100 have a high rate. Furthermore, compounds with higher complexities, in the neighborhood of 200, also show an appreciable probability of hitting in phenotypic screens. These compounds have inhibition data that are greater than 10 µM and are less active. This observation supports the analysis presented in this paper and exemplified in Fig. 4 indicating that screening libraries for phenotypic screens would benefit from the inclusion of molecules with higher degrees of complexity.

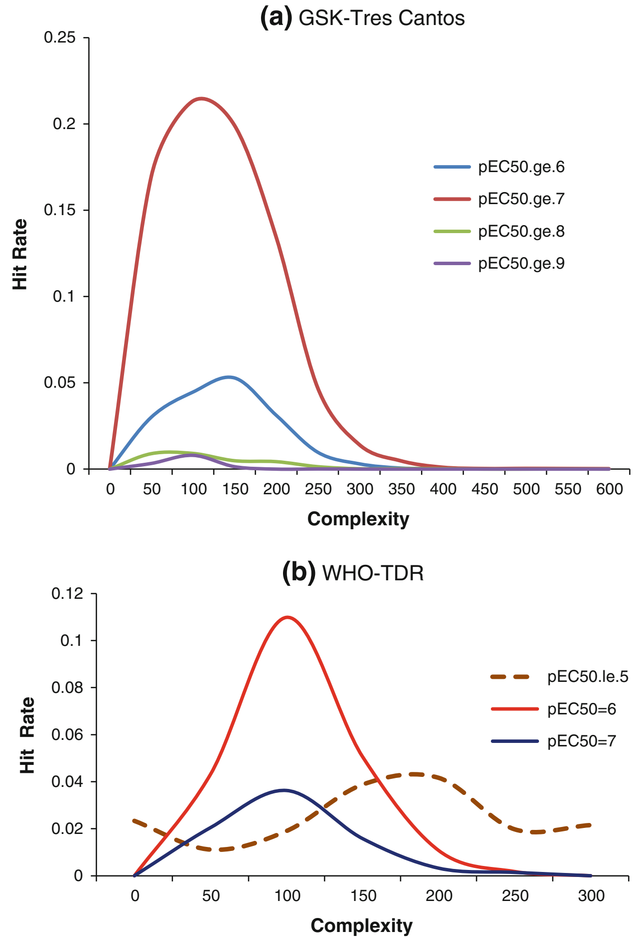

**Fig. 6** Analysis of hit rates with molecular complexity for phenotypic screens: **a** GSK-TresCantos Malaria *P. Falciparum* and **b** WHO–TDR: HAT, Chagas and Leishmaniasis

Modulating protein–protein interactions with small molecule inhibitors has received considerable attention [41–45]. Based on the results in Fig. 4, there is a finite probability of measuring the interactions if molecules of lower complexity are utilized to interrupt the intermolecular interactions at protein interfaces. However, a simple description of molecular recognition as used in this work does not include entropy effects and the geometry of the landscapes of the proteins. Clearly, while further interpretation of the model in these areas should be viewed with caution, the work presented here appears to support augmenting a collection of low complexity compounds with molecules of medium to higher complexity for high throughput phenotypic screens.

## Conclusions

In this paper, an extension to the Hann model of molecular complexity has been presented. The extension includes reverse and wrap-around modes of binding of the ligand to the protein and allows for partial interactions between the interacting systems. The estimation of the sensitivity of the measuring process has been evaluated using a functional form of the cumulative distribution function (CDF) of the Boltzmann distribution. As the CDF is dependent on the complexity of the ligand and the number of allowed interaction points **m** (within the polar representation) between the protein and the ligand, our extension establishes a link of the Hann model to the statistical nature of intermolecular recognition processes. Based on the results of the combinatorial evaluation of protein–ligand interactions within a one-dimensional model, the probability of observing hits based on ligand complexity, the selectivity and sensitivity of the measuring process has been rationalized. Explanations for the low hit rates from screening of corporate collections in HTS campaigns have been presented and suggestions have also been made for library design of screening collections in terms of the complexity of the compounds that are believed will increase the hit rate in screening experiments. Analysis of four datasets from ChemBL profiling hit rates with a measure for molecular complexity indicate that the augmentation of screening libraries with molecules of higher complexity may benefit the finding of hits (albeit weaker) for phenotypic screens.

**Acknowledgements** Shahul Nilar thanks Dr. Richard Lewis of the Novartis Institute for Biomedical Research, Basel, Switzerland and Dr. Peter Gedeck of the Novartis Institute for Tropical Diseases, Singapore for critically reading the manuscript and providing helpful suggestions. Dr. Ivica Res's help with some of the programming aspects of this work is gratefully acknowledged.

## References

1. Murray CW, Verdonk ML, Rees DC (2012) Experiences in fragment-based drug discovery. Trends in Pharmacological Science 33(5):224–232
2. Wyss DF, Wang Y-S, Eaton HL, Strickland C, Voigt JH, Zhu Z, Stamford AW (2012) Combining NMR and X-ray crystallography in fragment-based drug discovery: discovery of highly potent and selective BACE-1 inhibitors. Top Curr Chem 317:83–114
3. Guiguemde WA, Shelat AA, Garcia-Bustos JF, Diagana TT, Gamo F-J, Guy RK (2012) Global phenotypic screening for antimalarials. Chem Biol 19(1):116–129
4. Coxon GD, Cooper CB, Gillespie SH, McHugh TD (2012) Strategies and challenges involved in the discovery of new chemical entities during early-stage tuberculosis drug discovery. J Infect Dis 205(2):S258–S264
5. Payne DJ, Gwynn MN, Holmes DJ, Pompliano DL (2007) Drugs for bad bugs: confronting the challenges of antibacterial discovery. Nature Rev. Drug Disc. 6(1):29–40
6. Hajduk PJ, Galloway WRJD, Spring DR (2011) Drug discovery: a question of library design. Nature 470:42–43
7. Leach AR, Hann MM (2011) Molecular complexity and fragment-based drug discovery: ten years on. Curr Opin Chem Biol 15(4):489–496
8. Hann MM, Keserü GM (2012) Finding the sweet spot: the role of nature and nurture in medicinal chemistry. Nat Rev Drug Disc 11(5):355–365
9. Leeds JA, Schmitt EK, Krastel P (2006) Recent developments in antibacterial drug discovery: microbe-derived natural products—from collection to the clinic. Exp. Opin. Inv. Drugs. 15(3):211–226
10. Shoemaker RH, Scudiero DA, Melillo G, Currens MJ, Monks AP, Rabow AA, Covell DG, Sausville EA (2002) Application of high-throughput, molecular-targeted screening to anticancer drug discovery. Curr Topics Med Chem 2(3):229–246
11. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv. Drug. Del. Rev. 23(1–3):3–25
12. Oprea TI, Davis AM, Teague SJ, Leeson PD (2001) Is there a difference between leads and drugs? A historical perspective. J Chem Inf Comput Sci 41(5):1308–1315
13. We use of the term "promiscuous" to describe unselective compounds that interact with a number of protein targets
14. Dimova D, Hu Y, Bajorath J (2012) Matched molecular pair analysis of small molecule microarray data identifies promiscuity cliffs and reveals molecular origins of extreme compound promiscuity. J Med Chem 55(22):10220–10228
15. Clardy J, Fischbach MA, Walsh CT (2006) New antibiotics from bacterial natural products. Nat Biotechnol 24(12):1541–1550
16. Selzer P, Roth H-J, Ertl P, Schuffenhauer A (2005) Complex molecules: do they add value? Curr. Op. Chem. Biol. 9(3):310–316
17. Clemons PA, Wilson JA, Dančíka V, Muller S, Carrinski HA, Wagner BK, Koehler AN, Schreiber SL (2011) Quantifying structure and performance diversity for sets of small molecules comprising small-molecule screening collections. Proc Nat Acad Sci 108(17):6817–6822
18. Balaban AT, Mills D, Kodali V, Basak SC (2006) Complexity of chemical graphs in terms of size, branching, and cyclicity. SAR QSAR Environ Res 17(4):429–450
19. Nikolić S, Trinajstić N, Tolić IV (2000) Complexity of molecules. J Chem Inf Comput Sci 40(4):920–926
20. Hann MM, Leach AR, Harper G (2001) Molecular complexity and its impact on the probability of finding leads for drug discovery. J Chem Inf Comput Sci 41(3):856–864
21. Taylor NR, Cleasby A, Singh O, Skarzynski T, Wonacott AJ, Smith PW, Sollis SL, Howes PD, Cherry PC, Bethell R, Colman P, Varghese J (1998) Dihydropyrancarboxamides related to zanamivir: a new series of inhibitors of influenza virus sialidases. 2. Crystallographic and molecular modeling study of complexes of 4-amino-4H-pyran-6-carboxamides and sialidase from influenza virus types A and B. J Med Chem 41(6):798–807
22. Kitov PI, Sadowska JM, Mulvey G, Armstrong GD, Ling H, Pannu NS, Read RJ, Bundle DR (2000) Shiga-like toxins are neutralized by tailored multivalent carbohydrate ligands. Nature 403:669–672
23. Sneader W (1996) Drug prototypes and their exploitation. Wiley, New Jersey
24. Schuffenhauer A, Ruedisser S, Marzinzik AL, Jahnke W, Blommers M, Selzer P, Jacoby E (2005) Library design for fragment based screening. Curr Topics Med Chem 5(8):751–762
25. Jacoby E, Davies J, Blommers MJJ (2003) Design of small molecule libraries for NMR screening and other applications in drug discovery. Curr Top Med Chem 3:11–23
26. Klekota J, Roth FP (2008) Chemical substructures that enrich for biological activity. Bioinformatics 24(21):2518–2525. doi:10.1093/bioinformatics/btn479
27. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP (2011) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40(D1):D1100–D1107
28. Gamo F-L, Sanz LM, Vidal J, de Cozar C, Alvarez E, Lavandera J-L, Vanderwall DE, Green DVS, Kumar V, Hasan S, Brown JR, Peishoff CE, Cardon LR, Garcia-Bustos JF (2010) Thousands of chemical starting points for antimalarial lead identification. Nature 465(7296):305–310
29. Nwaka S, Besson D, Ramirez B, Maes L, Matheeussen A, Bickle Q, Mansour NR, Yousif F, Townson S, Gokool S, Cho-Ngwa F, Samje M, Misra-Bhattacharya S, Murthy PK, Fakorede F, Paris J-M, Yeates C, Ridley R, Van Voorhis WC, Geary T (2011) PLoS Negl. Trop. Dis. 5(12):e1412
30. Guner OF (2000) Pharmacophore perception, development and use in drug design. International University Line, La Jolla
31. Gozalbes R, Mosulén S, Carbajo RJ, Pineda-Lucena A (2009) Development and NMR validation of minimal pharmacophore hypotheses for the generation of fragment libraries enriched in heparanase inhibitors. J Comput Aided Mol Des 23(8):555–569
32. Leach AR, Gillet VJ, Lewis RA, Taylor R (2010) Three dimensional pharmacophore methods in drug discovery. J Med Chem 53(2):539–558
33. Ciulli A, Williams G, Smith AG, Blundell TL, Abell C (2006) Probing hot spots at protein–ligand binding sites: a fragment-based approach using biophysical methods. J Med Chem 49(16):4992–5000
34. Howard S, Berdini V, Boulstridge JA, Carr MG, Cross DM, Curry J, Devine LA, Early TR, Fazal L, Gill AL, Heathcote M, Maman S, Matthews JE, McMenamin RL, Navarro EF, O'Brien MA, O'Reilly M, Rees DC, Reule M, Tisi D, Williams G, Vinkovi M, Wyatt PG (2009) Fragment-based discovery of the pyrazol-4-yl urea (AT9283), a multitargeted kinase inhibitor with potent aurora kinase activity. J Med Chem 52(2):379–388
35. Jahnke W, Erlanson DE (2006) Fragment-based approaches in drug discovery. Wiley, New Jersey

36. Murray CW, Rees DC (2009) The rise of fragment-based drug discovery. Nature Chem 1(3):187–192
37. Baker M (2013) Fragment based drug discovery grows up. Nat Rev Drug Disc 12(1):5–7
38. Smith A (2002) Screening for drug discovery: the leading question. Nature 418(6896):453–459
39. Hert J, Irwin JJ, Laggner C, Keiser MJ, Shoichet BK (2009) Quantifying biogenic bias in screening libraries. Nature Chem. Biol. 5(7):479–483
40. Albert JS, Blomberg N, Breeze AL, Brown AJ, Burrows JN, Edwards PD, Folmer RH, Geschwindner S, Griffen EJ, Kenny PW, Nowak T, Olsson LL, Sanganee H, Shapiro AB (2007) Curr Top Med Chem 7:1600
41. Jubb H, Higueruelo AP, Winter A, Blundell TL (2012) Structural biology and drug discovery for protein–protein interactions. Trends Pharmacol Sci 33(5):241–248
42. Grimme D, Gonzalez-Ruiz D, Gohlke H (2012) Computational strategies and challenges for targeting protein–protein interactions with small molecules. Physico-chemical and Computational Approaches to Drug Discovery. London, UK, Royal Society of Chemistry (2012)
43. Arkin MR, Wells JA (2004) Small-molecule inhibitors of protein–protein interactions: progressing towards the dream. Nature Rev Drug Disc 3:301–317
44. Han S, Yin S, Hong YH, Mouhat S, Qiu S, Cao Z, Sabatier J-M, Wu Y, Li W (2010) Protein–protein recognition control by modulating electrostatic interactions. J Proteome Res 9(6):3118–3125
45. Mullard A (2012) Protein–protein interaction inhibitors get into the groove. Nat. Rev. Drug Disc. 11(3):173–175