ELSEVIER

www.elsevier.nl/locate/farmac

Il Farmaco 56 (2001) 137–139

IL FARMACO

Molecular diversity: how we measure it? Has it lived up to its promise?

Yvonne C. Martin *

Abbott Laboratories, D-47E AP10/2, 100 Abbott Park Rd., Abbott Park, IL 60064, USA

Received 30 November 2000

Abstract

Molecular diversity needs to be considered when designing a screening set, whether a set of individual compounds or a combinatorial library. We have found that traditional substructure descriptors encode information relative to the biological properties of molecules. Of the methods tested, Ward's clustering is most effective. Combinatorial library design also requires consideration of the method of deconvolution and the combinatorial constraint. We have developed a genetic algorithm solution to this problem. Lastly, predicting affinity of molecules for a macromolecular target has been addressed with the use of a potential of mean force. There continues to be an opportunity for innovative computational approaches to solving problems in medicinal chemistry. © 2001 Elsevier Science S.A. All rights reserved.

Keywords: Molecular diversity; Combinatorial library design; Deconvolution

1. Introduction

We care about molecular diversity because it is essential to our search for new and better drugs that treat currently untreatable diseases or that show improved safety, efficacy, or cost compared to existing drugs. These considerations have become important because of the opportunities presented by genomics, high throughput screening, and automated synthesis. In the case of genomics the challenge is to somehow find an active compound or a lead for synthesis to help validate a target; in the case of high throughput screening the challenge is to select the compounds for screening, usually to supplement a company's internal compound collection; in the case of automated chemistry the challenge is to design a library that will contain some active compounds. In all cases there is an intimate relationship between experimental and computational methods. Both the experimental and computational methods have expanded greatly in recent years.

High throughput screening requires diverse structures. Typical questions to be asked are: Is the company collection diverse enough that we would expect to find hits in most screening? Are the compounds offered for purchase or suggested as a universal library different enough from each other and the currently available compounds [1,2]? The problem is how to translate these questions into a computational procedure. A typical strategy is to group similar compounds together using cluster analysis and then to select compounds from different clusters as being diverse.

2. Designing the clustering strategy

The first issue in this strategy is how to represent the structures in the computer. Because our goal is to select compounds for biological testing we must select molecular descriptors that are relevant to biological testing. Medicinal chemistry experience has shown that molecules that appear diverse in the 2D-structure diagram as considered by the synthetic chemist can appear similar in three-dimensional properties to a biological receptor. This insight in turn can lead to a new series that matches the 3D requirements of the receptor but presents the possibility for a new patent. From the viewpoint of molecular diversity, the 2D substructure

* Corresponding author.
E-mail address: yvonne.c.martin@abbott.com (Y.C. Martin).

PII: S0014-827X(01)01010-2

encodes information about synthesis and patent coverage of the molecule, so increasing diversity in a screening database should increase the chance that an easily synthesized and patentable compound will be found. From the viewpoint of molecular diversity, the 3D structure encodes information about the steric and electrostatic properties of the molecule, so increasing the 3D diversity in a database should lead to hits in more screens.

The second issue in this strategy is what clustering method provides the best grouping of similar compounds together. Early in our studies we showed that compounds that are similar to each other in terms of their 2D-structure diagram are likely to have similar biological properties. We found that if a compound is 0.85 similar to an active compound it has an 80% chance of itself being active [3]. We first used these 2D descriptors to compare different clustering methods. We found that Ward's method is most successful at putting active compounds into clusters with other active compounds [4].

We then turned out attention to the molecular descriptors used in the clustering. We compared three different ways of encoding the 2D structure and three different methods of encoding the 3D-structure information. For the latter we wrote an expert system that generates a 3D structure, recognizes potentially charged and hydrogen-bonding groups, and calculates the distances between the relevant atoms. We decided to use a single conformation for this calculation. To our surprise the best performing descriptors were simple counts of substructures as used in structure searching systems [4].

Why did the substructures perform better than the other descriptors? We decided to repeat the clustering exercise but rather than use biological activity as the endpoint use measured and calculated physical properties. In particular, we examined the ability of other members of a cluster to predict measured octanol–water and cyclohexane–water $\log P$ and pK_a and to predict a number of calculated properties that encode molecular flexibility, shape, and hydrogen-bonding character. We observed the same order of descriptors and clustering methods as for biological activity. From this we conclude that the sub-structural descriptors contain information about physical and 3D properties in some useful balance for predicting biological activity [5].

3. Results of the clustering

Our early work showed that increasing the number of clusters in a database also increases the number of clusters in the hits [1]. We also found that datasets that are intuitively more diverse have fewer compounds per cluster than do datasets that contain many analogues. In fact, one screening set selected by hand by a medicinal chemist had 1.7 compounds per cluster whereas one combinatorial library had ten times as many compounds per cluster.

Often part of the high throughput of HTS comes from mixing compounds. The Abbott screening strategy places each compound in two orthogonal mixtures. A key to the successful execution of this strategy is that there not be similar compounds in any mixture — by using clustering to design the layout of the compounds we enabled the screening strategy.

An early strategy to select precursors for combinatorial libraries involved using a modification of our 3D features program described above. We added a special atom type for the position of attachment to the core. When we grouped precursors according to this method and selected one compound from each group, we found that the median 2D similarity of the precursors to the most similar compound was 0.6 whereas it was 0.85 in the original precursor list. Again we have confirmed that 2D and 3D structural descriptors contain overlapping information.

4. Genetic algorithm design of combinatorial libraries

In our combinatorial chemistry program we used mass spectrometry to deconvolute the structures, after release from the active bead. However, if many of the compounds in the library had the same mass, this would require individual synthesis of many molecules to identify the active one. At the same time, we required that the libraries be structurally diverse. How could we optimize diversity while simultaneously minimizing redundancy in the mass spectrum and obeying the combinatorial constraint? The question could be generalized to optimize the library with respect to structural diversity while constraining it to the combinatorial constraint and specified ranges of molecular weight and physical properties.

For example, consider the desire to design a 100×100 library with reagents selected from lists of 260 and 369 precursors. At first glance this problem seems simple. For this problem there are 93 240 possible products. However, there are 2.5×10^{169} possible ways to select the 10 000 compounds! Clearly, some sort of computer optimization is required.

We used a genetic algorithm to solve this problem [6]. The use or non-use of a particular precursor is encoded in a bit-string used as the chromosome. In a genetic algorithm potential solutions, libraries in this case, are generated at random for the first generation. Each solution is scored for its match to the target property, diversity and lack of mass spectrum redundancy in this case. The best scoring solutions are used

to form the next generation by a combination of crossover and mutations. The process proceeds through many generations until a suitable solution is found or until no improvement can be seen. For the example the best solution in the first generation had one mass redundancy of 80 and it used 69% of the possible clusters; at convergence the best solution had no mass redundancies of greater than 4 and diversity had increased to 79% of the available clusters.

5. Predicting potency from 3D structures of targets

Of course molecular diversity is of no use if the compounds selected are not active. While remaining useful for optimizing structures within a series, traditional and 3D-QSAR methods do not directly address diverse molecular structures. The promise is that knowledge of the three-dimensional structure of the biological target will increase the chance to discover active agents of diverse structure. However, predicting affinity of ligands to a protein target has remained a challenge.

The great increase in the number of experimentally determined 3D structures of protein ligand complexes provided us with an opportunity to investigate a new approach to this problem [7]. Specifically, we tabulated from protein ligand complexes the observed frequency of specific atom–atom contacts at different distances. From these frequency patterns we developed a potential of mean force that describes the free energy contribution of a particular atom-pair to binding affinity. Our approach includes desolvation terms implicitly. This potential of mean force (PMF) can be used both to determine the optimum placement of a ligand in a protein active site and to predict the relative affinity of different ligands to a protein. Note that it includes no fitting of the strength of observed protein–ligand complexes and that one function applies to all complexes.

In six of seven sets of ligand–protein complexes the PMF predicted relative affinity more accurately than LUDI [8], until then the best-predicting algorithm for this purpose [7].

The PMF was added to DOCK [9] and used in a docking study of 3247 compounds studied by SAR by NMR [10] of which 29 bind to the protein. The Amber scoring function enriched the actives toward higher scores by 1.54X whereas the PMF enriched the actives 2.0× [11]. The enrichment is especially pronounced with low-molecular weight compounds, rising to 2.77 (Amber enrichment doesn't change).

The PMF strategy for developing scoring functions has a promising start and will become even more accurate as workers analyze the atom-pair distance distributions of more and more protein–ligand complexes.

6. Conclusions

Molecular diversity considerations have led to rational strategies for designing screening libraries, whether they be selected from available compounds or combinatorial libraries. Molecular diversity is just one part of the equation: for whatever purpose the compounds are selected, considerations have to be made of chemical stability, avoidance of unwanted functional groups, physical properties, etc. With combinatorial libraries these considerations also include the deconvolution method. If the structure of the target is known, then the diversity also has to include forecast affinities of the compounds for the target. Advances in computational techniques have addressed each of these questions, but more insights are needed.

References

[1] Y.C. Martin, R.D. Brown, M.G. Bures, Quantifying diversity, in: J.P. Kerwin, E.M. Gordon (Eds.), Combinatorial Chemistry and Molecular Diversity in Drug Discovery, Wiley, New York, 1998, pp. 369–385.
[2] M.G. Bures, Y.C. Manin, Computational methods in molecular diversity and combinatorial chemistry, Curr. Opin. Chem. Biol. 2 (1998) 376–380.
[3] Y.C. Martin, M.G. Bures, R.D. Brown, Validated descriptors for diversity measurements and optimization, Pharm. Pharmacol. Commun. 4 (1998) 147–152.
[4] R.D. Brown, Y.G. Martin, Use of structure–activity data to compare structure-based clustering methods and descriptors for use in compound selection, J. Chem. Inf. Comput. Sci. 36 (1996) 572–584.
[5] R.D. Brown, Y.C. Martin, The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding, J. Chem. Inf. Comput. Sci. 37 (1997) 1–9.
[6] R.D. Brown, Y.C. Martin, Designing combinatorial library mixtures using a genetic algorithm, J. Med. Chem. 40 (1997) 2304–2313.
[7] I. Muegge, Y.C. Martin, A general and past scoring function for protein–ligand interactions: a simplified potential approach, J. Med. Chem. 42 (1999) 791–804.
[8] H.J. Böhm, The computer program Ludi: a new method for the de novo design of enzyme inhibitors, J. Comput.-Aided Mol. Design 6 (1992) 61–78.
[9] I.D. Kuntz, J.M. Blaney, S.J. Oatley, R. Langridge, T.J. Femn, A geometric approach to macromolecule–ligand interactions, J. Mol. Biol. 161 (1982) 269–288.
[10] S Shuker, P. Hajduk, R. Meadows, S. Pesik, Discovering high-affinity ligands for proteins: SAR by NMR, Science 274 (1996) 1531–1534.
[11] I. Muegge, Y.C. Martin, P.J. Hajduk, S.W. Fesik, Evaluation of PMF scoring in docking weak ligands to the Fk506 binding protein, J. Med. Chem. 42 (1999) 2498–2503.