Statistical Validation of Compound Structure-Selectivity Relationship Using NCI Cancer Database

Jing Li [1] **Chris Benner** [2] **Tao Jiang** [3] **Yingyao Zhou** [4]

Keywords: structure-selectivity relationship, structure-activity relationship, compound fingerprint, growth inhibition, cancer cell line.

Abstract

Existence of compound structure-activity relationship (SAR) and structure-selectivity relationship (SSR) are the fundamental hypotheses medicinal chemists rely on for lead selection and optimization during drug development. Although SAR has recently been statistically examined, previous studies did not cover SSR, an equally important assumption., For the current study, structural similarities (or distances) between compounds are defined by fingerprint-based Tanimoto distances, while selectivity similarities (or distances) are measured by the correlation between corresponding inhibition profiles available in the U.S. National Cancer Institute (NCI). Under these definitions, the validity of SSR assumption is examined in a statistically rigorous manner. Quantitative SSR similarity thresholds are derived in a tabular form, which in turn can be used to control the false positive rate of the SSR search. As an application of our statistical results, 178 seed compounds with good efficacy and selectivity against 60 cancer cell lines were identified and expanded into a larger set of 5469 potential novel anti-cancer compounds using the principle of SAR and SSR. The final collection of our 5469 compounds that are suspected to have strong inhibition against specific cancers based on our statistical study is web-accessible. This method is superior to random selection based approaches by an expected 2.6 fold enrichment. It can also be applied to discover novel compounds from other sources.

[1] Dept of Comp Sci, Univ of California, Riverside, CA 92521. E-mail: `jili@cs.ucr.edu` Tel: (909)787-2882 Fax: (909)787-4643. This author will be presenting the paper if accepted.

[2] Genomics Inst of the Novartis Res Foundation, 10675 John J. Hopkins Dr, San Diego CA 92121. E-mail: `cbenner@gnf.org` Tel: (858)812-1790 Fax: (858)812-1570

[3] Dept of Comp Sci, Univ of California, Riverside, CA 92521. E-mail: `jiang@cs.ucr.edu` Tel: (909)787-2991 Fax: (909)787-4643. Supported in part by NSF Grants CCR-9988353 and ITR-0085910, and National Key Project for Basic Research (973).

[4] Genomics Inst of the Novartis Res Foundation, 10675 John J. Hopkins Dr, San Diego CA 92121. E-mail: `yzhou@gnf.org` Tel: (858)812-1568 Fax: (858)812-1570

1 Introduction

High throughput screening technology has generated biological data for small molecules at unprecedented scale. Sieving through a million-compound collection and quickly identifying drug leads with both good efficacy and desired specificity/selectivity is a key step in modern drug discovery programs. During any drug discovery and development process, the belief that compounds of similar scaffolds also tend to share similar biological properties including both activity and selectivity is the cornerstone that medicinal chemists rely upon in order to identify and optimize drug leads [1]. Despite the fact that both SAR and SSR have been widely used, SAR was statistically studied only recently [1, 2] and SSR remains unexamined. Good efficacy reflected by SAR is a prerequisite for a drug; however, good specificity or selectivity reflected by SSR is usually more important. This is because a drug is designed to produce a very selective effect on a given pathology; any other drug activity possessed by that drug may lead to obfuscating side effects [3]. Instead of taking the validity of SSR for granted based on previous statistical studies on SAR, we applied t-test and χ^2 statistic to study the null hypothesis of SSR using NCI cancer database in this work.

Both structural and selectivity similarities (or distances) should be defined first for our purpose of study. It has become a common practice to measure structural similarity using binary fingerprint patterns, which allows the quick processing of large compound libraries. This method is also often used when more complicated pharmacophoric descriptors are not feasible to derive. Biological activities of compounds were previously either categorized in a binary (strong or weak) fashion [2] or simply characterized by their binding affinity [1]. Different from previous SAR studies, selectivity must be represented by a data vector consisting a biological profile across multiple assays instead. We then define selectivity similarity in terms of correlation of such data vectors and use it to answer the following two questions: 1) Is the binary fingerprint a valid structure descriptor in capturing biological selectivity? 2) What threshold should be applied during a structural selectivity search, in order to give the right false positive control on the resultant structure neighbors to our query compounds?

The public NCI anti-cancer drug discovery database contains about 250k compounds and 41k growth inhibition measurements (GI50) across 60 different cancer cell lines [4]. Such a collection provides a value data set for our statistical validation of SSR. In answering the first question, we applied t-test and χ^2 statistic to study the null hypothesis of SSR. Analyses were done using both a standard NCI compound set and the whole dataset. A conditional probability table was compiled to answer the second question, whose statistical soundness was validated by the γ statistic using all compound pairs with structural similarities greater than 0.5. We then applied these results to dataming the NCI collection for novel anti-cancer compounds. In constructing a good anti-cancer compound library, we first identified 11 compounds of high specificity and 167 compounds of high selectivity as seeds. Using structural similarity threshold of 0.8 as suggested by previous SAR study [1, 2] and our SSR conditional probability table, the SAR/SSR analysis yielded 5469 compounds of potentially good selectivity against a particular cancer cell line, among which 70% compounds do not have enough biological data at this time. Such a collection is valuable for pilot cancer study programs, as well as prioritizing future GI50 measurements.

2 Method and results

2.1 Data source and distance definition

As of March 2002, the NCI anti-cancer database contained 250k compounds for use in cancer drug development. A total of 41k compounds have GI50 measurements in at least one of the 60 cancer cell lines. Gene expression data for the 60-cancer cell lines are also publicly available and has been intensively studied to discover potential compound-gene networks [5]. For this work, standard cheminformatics fingerprint tools were used to define structural similarity between compounds. SMILES strings representing compound structures, downloaded from the NCI web site, were converted to binary fingerprints of 512 bits each using the GenerFP program from JChem (http://www.jchem.com). The binary bits capture compound structural fragments.

Tanimoto distance (Equation 1) was used to characterize similarities between two arbitrary compounds X and Y. GI50s for compounds across 60 cancer cell lines form an activity vector, from which a selectivity profile distance can be calculated using the Pearson correlation coefficient (Equation 2). Pairwise missing values were removed beforehand. Among the whole dataset, 123 compounds were considered as *standard* compounds because they have well known binding mechanisms [4, 6]. These are the targets of our first pilot study. We considered the complete dataset afterwards.

$$D_S(X,Y) = 1 - \frac{X \bigcap Y}{X \bigcup Y} \tag{1}$$

$$D_A(X,Y) = (1 - corr(X,Y))/2 \tag{2}$$

2.2 Preprocessing

The following filtering steps were applied to the compound collection. First, there are only about 36k compounds that have both structure and GI50 data available. Second, the GI50 vector of a compound should have a minimum variance of 0.015 and cannot have more than 20 identical elements. This is because GI50 for a particular cell line was assigned a default value if the compound was found inactive across all testing doses, which leads to round-off artifacts. About 12k compounds remained after this filtering step. Both structural and selectivity distances were then calculated pair wisely for our statistical analysis below.

2.3 Statistical analysis of the standard set

The scatter plot of both structural and selectivity distances in Figure 1 shows a trend of positive SSR correlation. A straightforward linear regression analysis gave a correlation coefficient of 0.226 ± 0.005 and a p-value of 10^{-16}. Although this confirms SSR as a non-trivial assumption, the linear model does not provide much prediction value due to its big residual error ($R^2 = 0.024$). To get results in an applicable form for medicinal chemists, conditional probabilities $Pr(SelectivityDistance \leq \alpha | StructureDistance \leq \beta)$ were estimated for different α and β as shown in Figure 2 (the tabular data are available at our web site). Using the result, a non-subjective structural distance threshold can be determined once a desired false positive rate is specified by a chemist.

2.4 Statistical analysis of the whole dataset

A histogram analysis indicated that the whole NCI collection may be over-represented by compounds with large structural distances ($D_S > 0.8$). Therefore, extra care has to be taken in the statistical analysis. Based on our preliminary pilot study using a jackknife-like re-sampling technique on the whole data, SSR for these dissimilar compound pairs is essentially non-significant. Considering the fact that similar compound pairs are the primary targets of interest to medicinal chemists, we focused on data points corresponding to structural distance less than 0.5 (chosen based on the γ statistic results described below) from now on. A similar conditional probability surface is shown as Figure 3 (see our web site for tabular data). Since this conditional probability table is compiled with a much larger compound set it might be less biased compared to the one based on the standard compound set.

We attempted to analyze quantitatively the positive correlation between structural and selectivity distances, which would establish the soundness of the above conditional probability table. As observed above, the linear model is not appropriate for validating SSR. One possible reason may lie in the heterogeneous nature of the compound structures in the collection. Therefore, we switched to a rank based non-parametric approach to access the statistical significance of our tabular results. Let D_S and D_A stand for the structural and selectivity distance variables, respectively. We first mapped the original continuous values into two new categorical ones D'_S and D'_A , i.e., group distance data into bins of width 0.1 ranging from 0 to 1. A 10×10

contingency table is constructed, where each cell holds the number of distance pairs (D_S, D_A) fallen into that category. The χ^2 statistic of the contingency table is 1.1×10^6 with 81 degrees of freedom and a p-value of 10^{-16}. A great dependency between the structural and the selectivity was again confirmed, but at a much larger scale this time.

The ordinal association of the (D'_S, D'_A) can be assessed by the γ statistic. Two data pairs (D'_{S1}, D'_{A1}) and (D'_{S2}, D'_{A2}) are defined as *concordant* if $D'_{S1} < D'_{S2}$ and $D'_{A1} < D'_{A2}$, or $D'_{S1} > D'_{S2}$ and $D'_{A1} > D'_{A2}$, and *discordant* otherwise. The numbers of concordant pairs and discordant pairs can be calculated using Equation 3, where Π_{ij} is an entry in the contingency table. The γ statistic defined in Equation 4 [7] is closer to 1 if the model provides more prediction value.

$$\Pi_c = 2\sum_i \sum_j \Pi_{ij} \big(\sum_{h>i} \sum_{k>j} \Pi_{hk}\big) \text{ and } \Pi_d = 2\sum_i \sum_j \Pi_{ij} \big(\sum_{h>i} \sum_{k<j} \Pi_{hk}\big) \tag{3}$$

$$\gamma = \frac{\Pi_c - \Pi_d}{\Pi_c + \Pi_d} \tag{4}$$

Distance cutoffs from 0.2 to 1.0 at an interval of 0.1 were used in the γ statistic, one at a time. Each cutoff was used to define the new boundary to simplify the original 10×10 contingency table into a 2×2 contingency table in order to find a proper structural distance threshold so that we could focus more on similar compounds. A structural distance threshold of 0.5 was chosen based on the γ values. The structural similar compounds were reexamined with different selectivity distance cutoffs, where γ ranged from 0.20 to 0.43. The result is quite satisfactory considering the fact that γ was near 0 when nonsimilar compounds were included.

2.5 Seed identification and novel compound discovery

To identify good seed compounds, a Z-score was defined for each compound as

$$Z = (GI50_{max} - mean_{remain})/stdv_{remain}.$$

There are a total of 5k compounds with Z-scores larger than 3, which means that these compounds inhibit a particular cancer cell line more strongly than others with 99.9% confidence. 11 compounds with high specificity were identified, which were selected by the criterion: $Z > 3$ and $mean_{remain} < 3$. The structures and GI50 heat-maps of these compounds are shown in Figures 4 and 5 (The heat-maps describe activities of compounds against some specific cancer cell lines). 167 compounds with high selectivity were identified, which were selected by the criterion: $Z > 3$ and $GI50_{max} - mean_{remain} > 3$. This corresponds to a 1000 fold average inhibition efficacy between the targeted cell line and others. By applying SAR/SSR analysis, a final collection of 5469 compounds were selected by searching the whole 250k compound collection using a 0.8 similarity cutoff in Tanimoto distance. Among this collection, only 1637 (30%) compounds have at lease one GI50 value and 880 (16%) compounds survived our preprocessing. The rest 70% would have been missed if a pure activity based approach were taken. Based on the conditional probability table, we expect 38% (2079) compounds to share the selectivity profiles with our seed compounds (correlation coefficient larger than 0.5). If a random selection procedure had been used to build a collection of the same size, the expectation of the true positive rate would have been only about 14.5% according to our sampling simulations. Therefore, our method has an expected 2.6 fold enrichment in terms of identifying compounds with desired selectivity profiles.

3 Conclusion

By introducing a new selectivity metric based on growth inhibition profiles across 60 cancer cell lines, we examined the validity of the structure-selectivity dependency assumption by applying t-test, χ^2 and γ statistics to the NCI cancer database. Besides capturing drug efficacy, the new selectivity metric emphasizes on drug specificity and selectivity. The quantitative conditional probability table, derived from the whole NCI compound collection, enables us to successfully identify a total number of 5469 potential anti-cancer compounds by SAR/SSR expansion of 178 seed compounds. Compared to a random selection approach, the current collection has expected 2.6 fold enrichment in identifying compounds with high selectivity (true positive rate of 38%). All results are publicly available from our web site (http://carrier.gnf.org/publications/NCISAR).

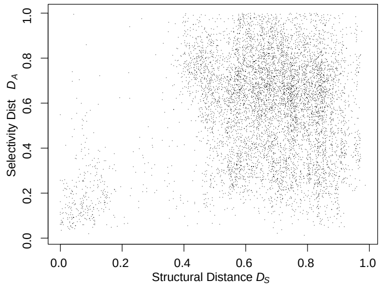

Figure 1

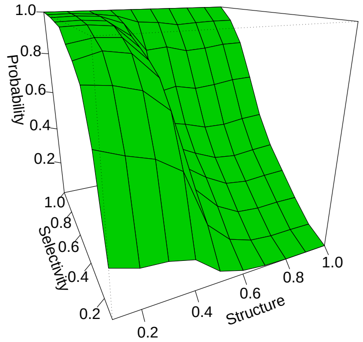

Figure 2

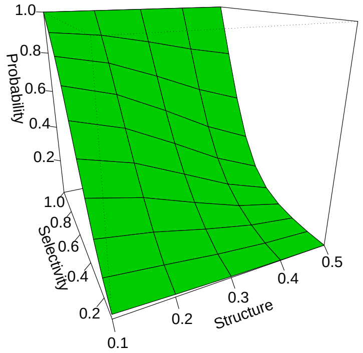

Figure 3

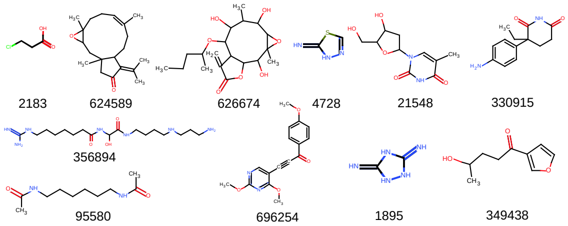

Figure 4

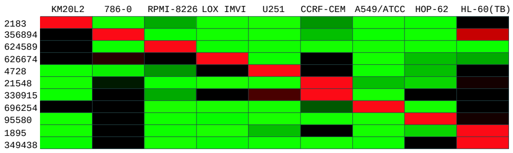

Figure 5

References

[1] Patterson D. E. *J. Med. Chem.*, 39(16):3049-3059, 1996

[2] Martin Y. C., Kofron J. L., Traphagen L. M. *J. Med. Chem.* 45(19):4350-8, 2002

[3] Lutz M. and Kenakin T., *Quantitative Molecular Pharmacology and Informatics in Drug Discovery*, John Wiley & Sons, LTD, 1999. page 25-27.

[4] Boyd, M. R. *Cancer: Principles and Practice of Oncology* 3:1-12, 1989.

[5] Scherf, U. et al. *Nature Genetics* 24:236-44, 2000

[6] Boyd, M. R. and Paul, K. D. *Drug Dev. Res.* 34:91-109, 1995

[7] Agresti,A. *Categorical Data Analysis*, 1990.