Bioorganic & Medicinal Chemistry Letters 20 (2010) 1338–1343

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry Letters

journal homepage: www.elsevier.com/locate/bmcl

The Privileged Chemical Space Predictor (PCSP): A computer program that identifies privileged chemical space from screens of modularly assembled chemical libraries

Steven J. Seedhouse, Lucas P. Labuda, Matthew D. Disney *

Department of Chemistry and The Center of Excellence in Bioinformatics and Life Sciences, University at Buffalo, The State University of New York, 657 Natural Sciences Complex, Buffalo, NY 14260, USA

ARTICLE INFO

Article history:
Received 23 November 2009
Revised 2 January 2010
Accepted 5 January 2010
Available online 11 January 2010

Keywords:
RNA
Chemical biology
High-throughput screening
Chemical microarrays
Computer program
Chemical space

ABSTRACT

Modularly assembled combinatorial libraries are often used to identify ligands that bind to and modulate the function of a protein or a nucleic acid. Much of the data from screening these compounds, however, is not efficiently utilized to define structure–activity relationships (SAR). If SAR data are accurately constructed, it can enable the design of more potent binders. Herein, we describe a computer program called Privileged Chemical Space Predictor (PCSP) that statistically determines SAR from high-throughput screening (HTS) data and then identifies features in small molecules that predispose them for binding a target. Features are scored for statistical significance and can be utilized to design improved second generation compounds or more target-focused libraries. The program's utility is demonstrated through analysis of a modularly assembled peptoid library that previously was screened for binding to and inhibiting a group I intron RNA from the fungal pathogen *Candida albicans*.

Combinatorial chemistry is commonly used to discover chemical genetics probes or therapeutics. The major advantage of a combinatorial approach is that it allows for the efficient synthesis and screening of large ligand libraries.[1] In synthesizing a large library, however, one must compromise between how focused the library is (target oriented synthesis, TOS) and how diverse it is (diversity oriented synthesis, DOS).[2–4] That is, does one test a narrowly focused landscape of chemical space such that the best ligand within this limited window is not overlooked or should one probe vast landscapes of chemical space in the hopes of identifying novel ligands? In the former case, novel binders are missed while in the latter case it is likely that the best inhibitor would not be identified. This conflict is inevitable but careful library design and analysis of screening data can partially mitigate this dilemma.

One approach that could merge these screening philosophies while also mitigating their drawbacks is to statistically analyze features in binders from a screen. This analysis would define privileged chemical space for a target and could be used to guide the rational design of second generation compounds. Such an approach is most easily applied to modularly assembled libraries of ligands such as peptoids, peptides or carbohydrates,[5–9] because each module and its relative positioning to other modules can be statistically analyzed to determine features that afford a potent ligand.

Small molecule–RNA interactions are one area where there is an underdeveloped knowledge base regarding the types of chemical ligands that are biased for binding a target,[10] although there continues to be progress in this area.[11,12] Previously, a modularly assembled library of RNA-focused peptoids was designed, synthesized, and tested via microarray for binding to an essential group I ribozyme from the pathogenic fungus *Candida albicans* (Fig. 1).[13] In that report, a library of peptoids was constructed with nine different monomers assembled across three positions; only 14% of the total possible number of library members were synthesized and tested. The compound library was screened by microarray[14] and manual statistical analysis of the binders identified features in ligands that predispose them for binding to and inhibiting the group I intron. This analysis allowed for the design of improved ligands that were not tested in the original screen but would have been members of a comprehensive library. Such manual analysis, however, is not practical with larger libraries.

In order to streamline analysis of SAR data, a computational approach was developed and is described herein in which HTS data are statistically analyzed to identify privileged RNA-binding space. The Windows-based program that implements this approach is called Privileged Chemical Space Predictor (PCSP). The output of the program is an analysis of the features in the compounds that

* Corresponding author.
E-mail address: mddisney@buffalo.edu (M.D. Disney).

0960-894X/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmcl.2010.01.017

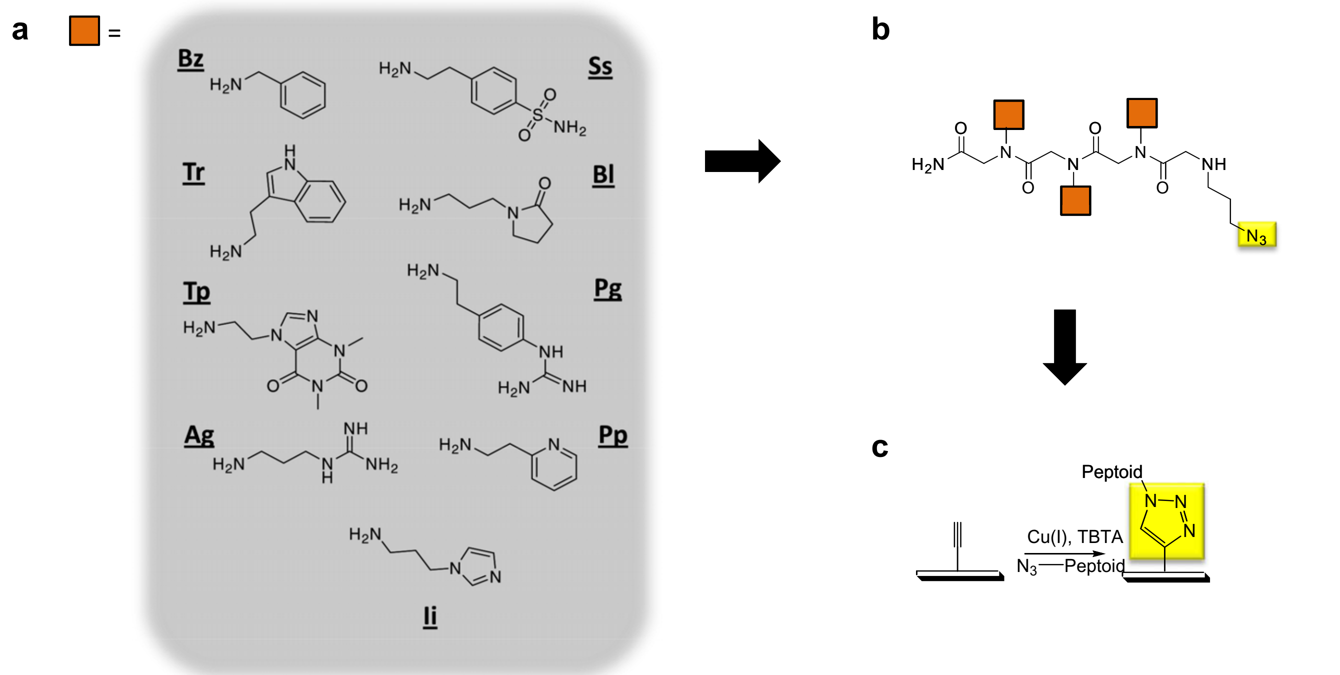


Figure 1. Example of a rational 'ground-up' approach towards library design: (a) building blocks are chosen and functionalized for combinatorial peptoid synthesis; the corresponding two-letter abbreviations are shown, (b) building blocks are modularly assembled onto a peptoid scaffold with an azide linker, (c) compounds are immobilized onto a microarray surface via the azide linker using 1,3 Huisgen dipolar cycloaddition reaction for high-throughput screening.

are predisposed for binding and inhibiting a target, or privileged chemical space. The PCSP program can be downloaded free of charge at http://www.nsm.buffalo.edu/Research/rna.

The user interface of PCSP was designed to provide sufficient control over data analysis and data output (Fig. 2). First, the user can define a two-letter nomenclature for each building block of a modularly assembled ligand. Examples using the peptoid library described above are shown in Figure 1A (**Bl**, **Ag**, **Tr**, etc.). Next, data are uploaded into PCSP from a standard text file that contains columns with the identity of each compound and its signal for binding and/or its potency for inhibiting a target. PCSP normalizes the binding or inhibition values and calculates statistical trends once a cut-off is assigned to score a ligand as a positive hit (Figs. 2 and 3). For example, in our studies using microarrays, we consider a ligand to be a binder if it gives ⩾20% of the signal relative to the best binder (greatest signal) on an array. PCSP then determines statistical trends by computing *Z*-scores that correspond to a confidence level >95%. *Z*-scores are calculated from Eq. 1:[15]

$$Z_{\text{obs}} = \frac{p_1 - p_2}{\sqrt{\phi(1-\phi)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \qquad \phi = (n_1p_1 + n_2p_2)/(n_1 + n_2) \tag{1}$$

where n_1 is the number of compounds containing the trend, n_2 is the number of compounds without the trend, p_1 is the proportion of binders amongst a population of compounds with a specific trend, p_2 is the proportion of binders amongst compounds that do not contain the trend, and ϕ is the pooled sample proportion.

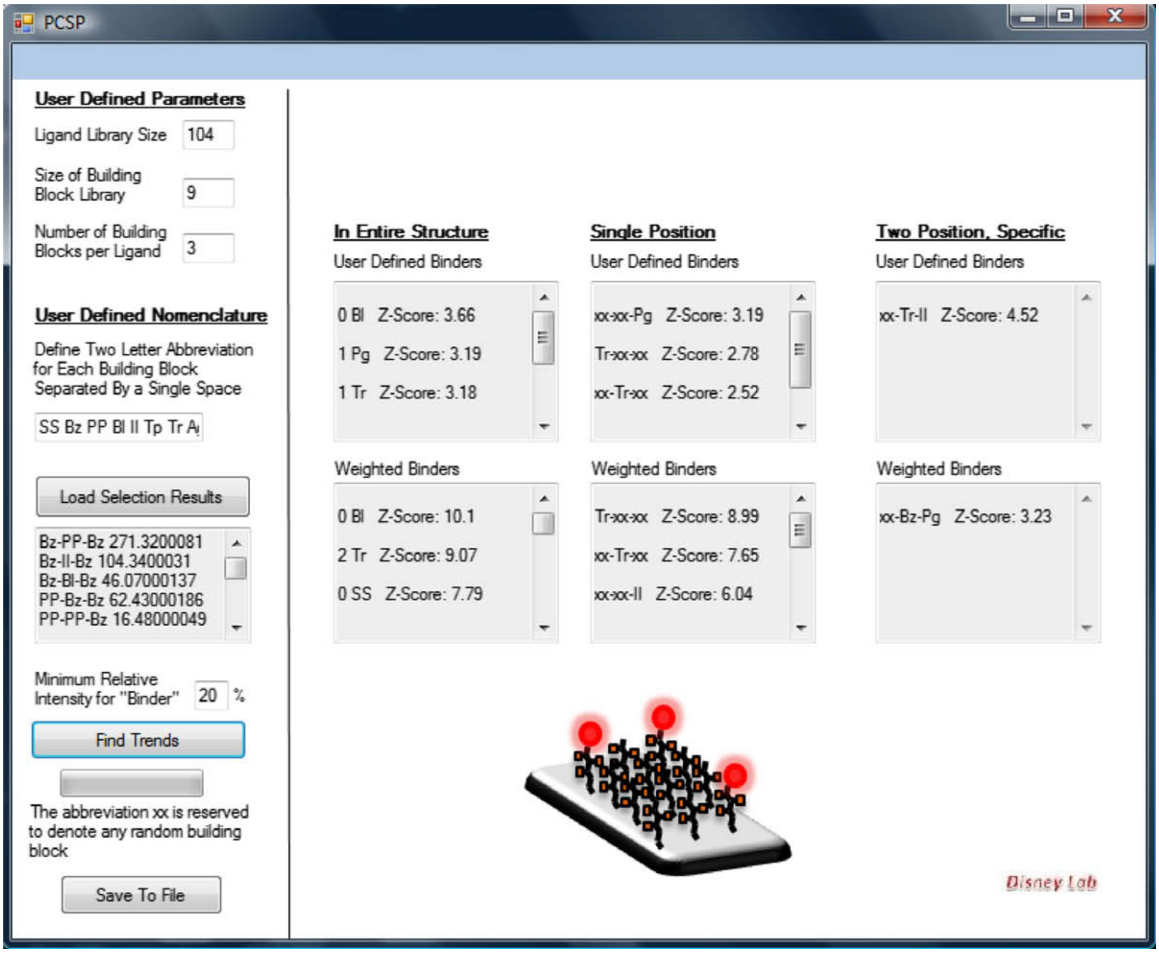
Z-scores are easily converted to *p*-values,[16] which provide a direct assessment of the probability that the associated trend represents privileged chemical space. For example, the null hypothesis for this system is that if specific building blocks and their relative arrangement on peptoids have no particular bias for binding to a target, then populations of binders and non-binders should be similar in composition. A two-tailed *p*-value represents the probability that the observed proportion of a specific trend could occur if in fact there is no such bias. Therefore, a trend with an associated two-tailed *p*-value of <0.05 statistically confers >95% confidence of rejecting the null hypothesis.

PCSP also provides weighted scores for each ligand binder (Figs. 2 and 3). This value is calculated using a novel formula (Eq. 2) that is analogous to that used to determine statistical *Z*-scores. In this weighted function, each ligand is normalized for its binding signal on an array or any other measured value (e.g., IC_{50}). In this way, both the populations that statistically represent binders and the relative affinity or potency of those binders are taken into consideration.

$$Z_{\text{weighted}} = \frac{\gamma_1 - \gamma_2}{\sqrt{\phi(1-\phi)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \qquad \phi = (n_1\gamma_1 + n_2\gamma_2/n_1 + n_2) \tag{2}$$

where n_1 is the number of compounds containing the trend, n_2 is the number of compounds without the trend, γ_1 is the average relative binding signal amongst a population of compounds with a specific trend, γ_2 is the average relative binding signal amongst those which do not contain that trend, and ϕ is the pooled sample proportion.

To validate the effectiveness of PCSP, data from the previously reported microarray screen of modularly assembled ligands and the *C. albicans* group I intron[13] were loaded into PCSP and analyzed. Based on the raw microarray signals, PCSP identified numerous statistically significant trends (>95% confidence level) for composition of the ligands. A schematic of the analysis performed by PCSP is shown in Figure 3. Three classes of trends are revealed by PCSP that describe chemical space that predisposes a compound for binding. These classes include the occurrence or absence of particular modules, the position of each module, and the positioning of modules relative to each other.



PCSP

User Defined Parameters

Ligand Library Size 104

Size of Building Block Library 9

Number of Building Blocks per Ligand 3

User Defined Nomenclature

Define Two Letter Abbreviation for Each Building Block Separated By a Single Space

SS Bz PP Bl Il Tp Tr A

```
Bz-PP-Bz 271.3200081
Bz-Il-Bz 104.3400031
Bz-Bl-Bz 46.07000137
PP-Bz-Bz 62.43000186
PP-PP-Bz 16.48000049
```

Minimum Relative Intensity for "Binder" 20 %

The abbreviation xx is reserved to denote any random building block

In Entire Structure

User Defined Binders

```
0 Bl  Z-Score: 3.66
1 Pg  Z-Score: 3.19
1 Tr  Z-Score: 3.18
```

Weighted Binders

```
0 Bl  Z-Score: 10.1
2 Tr  Z-Score: 9.07
0 SS  Z-Score: 7.79
```

Single Position

User Defined Binders

```
xx-xx-Pg  Z-Score: 3.19
Tr-xx-xx  Z-Score: 2.78
xx-Tr-xx  Z-Score: 2.52
```

Weighted Binders

```
Tr-xx-xx  Z-Score: 8.99
xx-Tr-xx  Z-Score: 7.65
xx-xx-Il  Z-Score: 6.04
```

Two Position, Specific

User Defined Binders

```
xx-Tr-Il  Z-Score: 4.52
```

Weighted Binders

```
xx-Bz-Pg  Z-Score: 3.23
```

Figure 2. PCSP user interface showing the results from the analysis of an uploaded data file and a trend query.

All trends identified by PCSP with Z_{obs} corresponding to >95% confidence level and a $Z_{weighted}$ >2 are listed in Table 1. The most significant trends for building block composition included compounds that contain exactly one **Tr** (Z_{obs} = 3.18, two-tailed *p*-value = 0.0015), one **Pg** (Z_{obs} = 3.19, two-tailed *p*-value = 0.0014), or 0 **Bl** (Z_{obs} = 3.66, two-tailed *p*-value = 0.0003). Also, the weighted formula revealed that in addition to those peptoids containing 0 **Bl** building blocks, compounds containing two **Tr** building blocks displayed unusually high binding signals. Thus, for future design of peptoids targeting this RNA, the **Bl** building block should be avoided while the **Tr** building block should be included.

PCSP also determined that the positional dependence of certain building blocks were favorable with >95% confidence: **xx–xx–Pg** (Z_{obs} = 3.19, two-tailed *p*-value = 0.0014), **Tr–xx–xx** (Z_{obs} = 2.78, two-tailed *p*-value = 0.0054), and **xx–Tr–xx** (Z_{obs} = 2.52, *p* = 0.0117), where **xx** denotes any building block. The three most relevant weighted trends of this type were **Tr–xx–xx**, **xx–Tr–xx**, and **xx–xx–Ii**. In addition to the positioning of single building blocks, the positions of two ligand modules relative to one another are also analyzed by PCSP, and **xx–Tr–Ii** (Z_{obs} = 4.52, two-tailed *p*-value <0.0001) is statistically significant. It was predicted by the weighted formula that **xx–Bz–Pg** may also be a good combination of building blocks for high affinity binding.

In the previous report,[13] new compounds were synthesized based on manual statistical analysis of the screening data. Two second generation compounds, **Tr–Tr–Pg** and **Tr–Tp–Pg**, were designed by replacing the building block in the third position of potent first generation inhibitors (**Tr–Tr–Ii** and **Tr–Tp–Ii**) with **Pg** (Table 1, two-tailed *p*-value = 0.0014). Both **Tr–Tr–Pg** and **Tr–Tp–Pg** are more potent than the respective parent compound. Other second generation compounds with four points of diversity were synthesized using the results of statistical analysis, which determined with the highest confidence level that the presence of **Pg** and **Tr** building blocks were important features for binding to the *C. albicans* group I intron. Therefore, **Pg** was placed in the fourth variable position as a side chain element to already potent ligands. This addition provided the best inhibitors, with compound **Tr–Tr–Tr–Pg** being the most potent of all compounds studied ($\geqslant$5-fold more potent than all first generation inhibitors). Interestingly, the presence of two **Tr** building blocks had the highest $Z_{weighted}$ score (Table 1, 9.07). The presence of **Tr** in the first position has a $Z_{weighted}$ score of 8.99 while **Tr** in the second position has a $Z_{weighted}$ score of 7.65 (Table 1). (It should be noted that **Tr–Tr–Tr** constituted only one member of the peptoid library screened in the study. Therefore, statistical analysis cannot meaningfully assign Z_{obs} or $Z_{weighted}$ values. In contrast, multiple members of the library contained one or two **Tr** building blocks in different positions.) The presence of one **Pg** building block also represents privileged chemical space and has a $Z_{weighted}$ score of 4.80 (Table 1).

Taken together, the first and second generation structures and their relative potencies are in good agreement with the statistical analysis performed by PCSP (Fig. 4 and Table 1), indicating its potential utility in future RNA–ligand or protein–ligands screens of

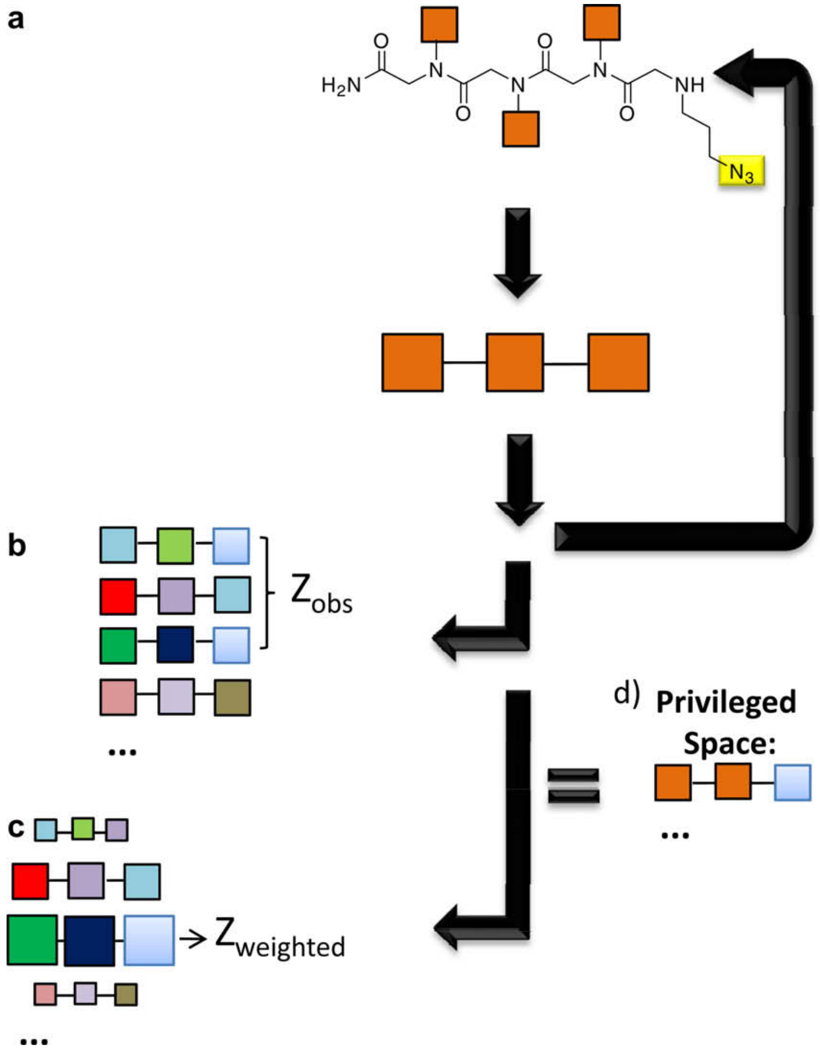


Figure 3. Schematic of PCSP function: (a) each ligand, described by its building blocks' abbreviations, is analyzed individually by PCSP, (b) the composition of each binder and non-binder is recorded for statistical analysis of the two populations according to their composition, (c) the composition of each ligand is scaled to its binding signal or affinity as determined by a high-throughput screen and used to calculate weighted scores for each trend, and (d) privileged chemical space is predicted according to the statistical trends found amongst binders and the weighted score.

Table 1
The privileged chemical space within a modularly assembled peptoid library identified for the *C. albicans* group I intron by PCSP[a]

By Z_{obs}	Normalized binding signal	IC_{50} (μM)	By $Z_{weighted}$	Normalized binding signal	IC_{50} (μM)
Number of building blocks					
0 **Bl** 3.66			0 **Bl** 10.10		
1 **Pg** 3.19	0.21–0.55	399–1733	2 **Tr** 9.07	1.0	158
1 **Tr** 3.18	0.32–0.76	150–817	0 **Ss** 7.79		
0 **Pp** 2.23			0 **Pp** 7.60		
			0 **Ag** 7.16		
			1 **Tr** 7.05	0.32–0.76	150–817
			0 **Bz** 6.61		
			1 **Pg** 4.80	0.21–0.55	399–1733
			0 **Tp** 4.67		
			1 **Ii** 4.43	0.32–1.0	158–1733
			2 **Ii** 4.34	0.21–0.76	157–2200
			1 **Tp** 3.98	0.32–0.61	150–2200
			0 **Pg** 3.65		
Position of individual building blocks					
xx–xx–Pg[b] 3.19	0.21–0.55	399–1733	**Tr–xx–xx**[b] 8.99	0.32–1.0	150–158
Tr–xx–xx[b] 2.78	0.32–1.0	150–158	**xx–Tr–xx**[b] 7.65	0.38–1.0	158–817
xx–Tr–xx[b] 2.52	0.38–1.0	158–817	**xx–xx–Ii**[b] 6.04	0.32–1.0	150–2200
			xx–xx–Pg[b] 4.74	0.21–0.55	399–1733
			Ii–xx–xx[b] 2.32	0.21–0.55	399–2200
Position of building blocks relative to each other					
xx–Tr–Ii[b] 4.52	0.38–1.0	158–817	**xx–Bz–Pg**[b] 3.23	0.55	—

[a] Only Z_{obs} corresponding to a >95% confidence level and $Z_{weighted}$ >2 are shown.
[b] **xx**″ denotes any building block.

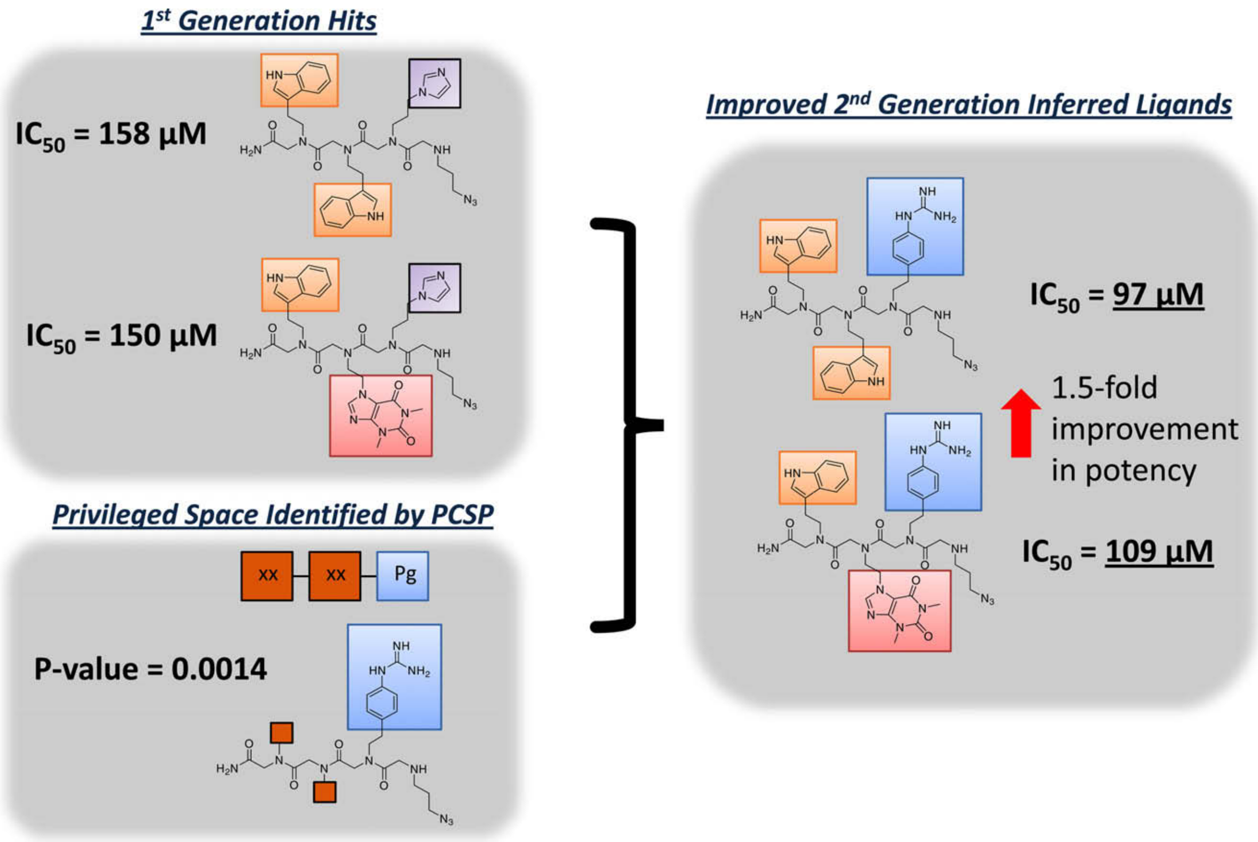

1st Generation Hits
IC50 = 158 µM
IC50 = 150 µM
Privileged Space Identified by PCSP
xx
xx
Pg
P-value = 0.0014
Improved 2nd Generation Inferred Ligands
IC50 = 97 µM
1.5-fold improvement in potency
IC50 = 109 µM
HN
NH
H2N
O
N
N3
NH2

Figure 4. Statistical interpretation from PCSP can be used to design second generation compounds with increased potency by modifying initial hits to contain the identified trends.

any scale. For example, one of the best first generation inhibitors, **Tr**–**Tr**–**Ii**, also has the highest binding signal on the array. This compound is composed of statistically significant trends with large Z_{obs} and $Z_{weighted}$ values. Two **Tr** building blocks corresponds to a $Z_{weighted}$ value of 9.07 while one **Ii** has a $Z_{weighted}$ value of 4.43. In addition, this compound contains trends that depend on the position of individual building blocks or the positions of the building blocks relative to each other. PCSP identified the following positional trends: **Tr**–**xx**–**xx** ($Z_{weighted} = 8.99$), **xx**–**Tr**–**xx** ($Z_{weighted} = 7.65$), **xx**–**xx**–**Ii** ($Z_{weighted} = 6.04$), and **xx**–**Tr**–**Ii** ($Z_{obs} = 4.52$). The program also identified chemical space that should be avoided. The absence of **Bl** building blocks has a $Z_{weighted}$ value of 10.10. Interestingly, **Tr**–**Tr**–**Bl** does not bind to the *C. albicans* group I ribozyme when displayed on the microarray surface.[13] Another compound that contains **Bl**, **Pp**–**Bl**–**Bz**, does not bind the group I ribozyme and is a poor inhibitor with an IC_{50} >5 mM. This compound has three features that are selected against according to their $Z_{weighted}$ values: 0 **Pp** building blocks has a $Z_{weighted}$ value of 7.60; 0 **Bl** building blocks has a $Z_{weighted}$ value of 10.10; and 0 **Bz** building blocks has a $Z_{weighted}$ value of 6.61. Therefore, PCSP successfully identified both positive and negative features for inhibition of the *C. albicans* ribozyme by statistical analysis of the binding signals from the microarray.

In summary, the PCSP program was developed to statistically analyze a modularly assembled library to identify features that predispose ligands for binding RNA. The program was able to quickly determine significant trends from a published microarray study on the binding of RNA-focused peptoids to a validated RNA target.[13] PCSP should also be generally useful for application to other biomolecular targets besides RNA. For example, modularly assembled peptoid libraries displayed on microarrays or beads have been screened to identify high affinity ligands that bind proteins.[17–19] Moreover, PCSP could be used to probe SAR data and define privileged diversity element space and the privileged relative positioning of diversity elements in any combinatorial library. The computational approach using PCSP expedites analysis of SAR data, can reveal statistically relevant trends that might otherwise be overlooked, and can be easily applied to screens of larger modularly assembled ligand libraries. Such studies should streamline analysis of HTS and SAR data to establish more information on the types of modularly assembled ligands that are biased for binding RNA or other targets that are present in the genome.

Acknowledgment

We thank the National Institutes of Health for funding this work (RO1-GM079235).

Supplementary data

Supplementary data (these data include the source code for the PSCP computer program) associated with this article can be found, in the online version, at doi:10.1016/j.bmcl.2010.01.017.

References and notes

1. Geysen, H. M.; Schoenen, F.; Wagner, D.; Wagner, R. *Nat. Rev. Drug Disc.* **2003**, *2*, 222.
2. Burke, M. D.; Schreiber, S. L. *Angew. Chem., Int. Ed.* **2004**, *43*, 46.
3. Fergus, S.; Bender, A.; Spring, D. R. *Curr. Opin. Chem. Biol.* **2005**, *9*, 304.
4. Schreiber, S. L. *Science* **2000**, *287*, 1964.
5. Simon, R. J.; Kania, R. S.; Zuckermann, R. N.; Huebner, V. D.; Jewell, D. A.; Banville, S.; Ng, S.; Wang, L.; Rosenberg, S.; Marlowe, C. K.; Spellmeyer, D. C.;

Tan, R. Y.; Frankel, A. D.; Santi, D. V.; Cohen, F. E.; Bartlett, P. A. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 9367.

6. Olivos, H. J.; Alluri, P. G.; Reddy, M. M.; Salony, D.; Kodadek, T. *Org. Lett.* **2002**, *4*, 4057.
7. Liu, Y.; Palma, A. S.; Feizi, T. *Biol. Chem.* **2009**, *390*, 647.
8. Ratner, D. M.; Adams, E. W.; Disney, M. D.; Seeberger, P. H. *ChemBiochem* **2004**, *5*, 1375.
9. Winkler, D. F.; Hilpert, K.; Brandt, O.; Hancock, R. E. *Methods Mol. Biol.* **2009**, *570*, 157.
10. Thomas, J. R.; Hergenrother, P. J. *Chem. Rev.* **2008**, *108*, 1171.
11. Parsons, J.; Castaldi, M. P.; Dutta, S.; Dibrov, S. M.; Wyles, D. L.; Hermann, T. *Nat. Chem. Biol.* **2009**, *5*, 823.
12. Marcheschi, R. J.; Mouzakis, K. D.; Butcher, S. E. *ACS Chem. Biol.* **2009**, *4*, 844.
13. Labuda, L. P.; Pushechnikov, A.; Disney, M. D. *ACS Chem. Biol.* **2009**, *4*, 299.
14. MacBeath, G.; Koehler, A. N.; Schreiber, S. L. *J. Am. Chem. Soc.* **1999**, *121*, 7967.
15. Weiss, N. A.; Hassett, M. J. *Introductory Statistics*; Addison-Wesley Pub. Co.: Reading, Mass., 1982.
16. To convert Z_{obs} to a two-tailed *p*-value, standard statistics tables can be used. For example, see Table 1 in 15.
17. Lim, H. S.; Reddy, M. M.; Xiao, X.; Wilson, J.; Wilson, R.; Connell, S.; Kodadek, T. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 3866.
18. Udugamasooriya, D. G.; Dineen, S. P.; Brekken, R. A.; Kodadek, T. *J. Am. Chem. Soc.* **2008**, *130*, 5744.
19. Zuckermann, R. N.; Kodadek, T. *Curr. Opin. Mol. Ther.* **2009**, *11*, 299.