BRIEF COMMUNICATION
PUBLISHED ONLINE: 18 DECEMBER 2011 | DOI: 10.1038/NCHEMBIO.732
nature chemical biology

Chemical informatics and target identification in a zebrafish phenotypic screen

Christian Laggner[1,8], David Kokel[2,3,8], Vincent Setola[4,5,8], Alexandra Tolia[6,7], Henry Lin[1], John J Irwin[1], Michael J Keiser[1], Chung Yan J Cheung[2,3], Daniel L Minor Jr[6,7], Bryan L Roth[4,5]*, Randall T Peterson[2,3]* & Brian K Shoichet[1]*

Target identification is a core challenge in chemical genetics. Here we use chemical similarity to computationally predict the targets of 586 compounds that were active in a zebrafish behavioral assay. Among 20 predictions tested, 11 compounds had activities ranging from 1 nM to 10,000 nM on the predicted targets. The roles of two of these targets were tested in the original zebrafish phenotype. Prediction of targets from chemotype is rapid and may be generally applicable.

Chemical genetics seeks to identify the targets responsible for phenotypes responding to organic molecules, just as genetic screens identify the molecular players involved in cellular processes. Because the chemical perturbation is nongenetic, targets must be identified biochemically[1] or proteomically[2,3] or by phenotypic pattern recognition[4,5]. Such approaches can be laborious, and many low-abundance proteins are outside of their purview. Several investigators thus have turned to inference-based methods that combine experiment and computation. Multidimensional screening of hundreds of cell lines by tens of thousands of molecules has led to patterns that can illuminate target identity[6,7], as can matching the transcriptional patterns provoked by an organic molecule to those provoked by molecules with known mechanisms[8,9].

Less explored are purely computational methods for target identification, especially those motivated by ligand structure, the basis of target identification in classical pharmacology[10]. Recently, large ligand-protein databases annotated hundreds of thousands of ligands to thousands of targets. This has allowed receptor relationships to be redrawn on the basis of shared ligands[11] and ligand-disease associations[12], which have been exploited to predict previously unknown targets for drugs[13,14]. By the same logic, such an approach might be used to predict the targets of organic molecules active in phenotypic screens.

We thus sought to identify the targets of 681 neuroactive molecules from a 14,000-compound phenotypic screen in zebrafish embryos by measuring the modulation of characteristic movements in response to a series of light flashes (a photomotor response, PMR)[15]. For each compound, patterns were observed for 8–10 embryos, and 14 behavioral features of the PMR in compound-treated animals were compared to the behaviors in untreated animals (**Supplementary Results**, **Supplementary Fig. 1** and **Supplementary Methods**)[15]. Of the 681 active molecules, 162 were drugs, chemical probes or naturally occurring molecules, 61 of which had one or more targets annotated in the ChEMBL database (https://www.ebi.ac.uk/chembldb/), leaving a total of 620 compounds unannotated. We computationally screened the 681 active compounds against a filtered version of ChEMBL that annotates over 167,000 organic molecules for activities against over 2,000 molecular targets using the similarity ensemble approach (SEA)[13,16]. SEA scores shared patterns of chemical functionality between the 'bait' molecules and all of the ligands annotated to a target (the ligand-target set) using one of several topological, bit-string fingerprints[17,18]. Similarity values are measured by Tanimoto coefficients (T_C)[19], which range from 0 (no bits in common) to 1 (all bits shared). The similarities between a bait molecule and a ligand-target set were summed and compared to those expected to occur at random. The BLAST statistical machinery was used to calculate expectation values (*E* values) for similarity versus a random background[13,16,20]. In a variation introduced here, bait molecules bearing formal charges that are atypical for a ligand-target set were excluded (**Supplementary Fig. 2**).

The method is limited to those targets with known ligands. Still, most target categories are covered, all of which are 'ligandable'. For instance, using the widely used ECFP4 fingerprint[17], 473 of the 681 molecules were predicted to be active on 945 targets with *E* values better than 10^{-5} (**Fig. 1**); at a more stringent *E* value of 10^{-20}, 284 molecules were predicted to be active on 404 targets. Similar results were obtained for the other fingerprints, resulting in combined

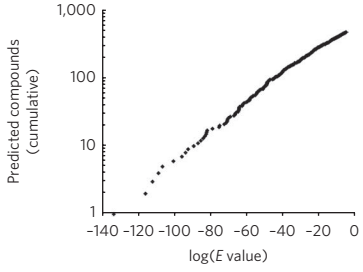

Figure 1 | Compounds for which at least one target is predicted. The number of predicted compounds versus the cutoff values for *E* values; higher *E* values are less stringent. Only those below (better than) 10^{-5} were considered in this analysis.

[1]Department of Pharmaceutical Chemistry, University of California, San Francisco, California, USA. [2]Cardiovascular Research Center and Division of Cardiology, Department of Medicine, Massachusetts General Hospital, Harvard Medical School, Charlestown, Massachusetts, USA. [3]Broad Institute, Cambridge, Massachusetts, USA. [4]Department of Pharmacology, University of North Carolina Chapel Hill School of Medicine, Chapel Hill, North Carolina, USA. [5]National Institute of Mental Health Psychoactive Drug Screening Program, University of North Carolina Chapel Hill School of Medicine, Chapel Hill, North Carolina, USA. [6]Cardiovascular Research Institute, University of California, San Francisco, California, USA. [7]Department of Biochemistry and Biophysics, University of California, San Francisco, California, USA. [8]These authors contributed equally to this work. *e-mail: bryan_roth@med.unc.edu or peterson@cvrc.mgh.harvard.edu or shoichet@cgl.ucsf.edu

predictions for 586 compounds with *E* values $<10^{-5}$. Targets with strong predictions included G protein–coupled receptors (GPCRs), ligand-gated ion channels, nuclear hormone receptors, transporters and soluble enzymes (**Supplementary Table 1**).

We emphasized those targets that were predicted with strong *E* values, were previously unknown for the phenotypically active molecules and were experimentally accessible to us. For this study, we investigated GPCRs, ion channels, transporters and kinases. Many target predictions were unknown in ChEMBL but were subsequently found through a literature search (**Supplementary Table 2**). Because we were interested in new predictions, we focused on 20 compounds predicted against unknown targets. Of these compounds, 11 were active at 22 of their predicted 31 targets (**Table 1**, **Supplementary Table 3** and **Supplementary Fig. 3**) and 9 compounds had no measurable activity at their predicted targets, and these predictions were considered falsified (**Supplementary Table 4**). Potencies, measured in full concentration-response, ranged from about 10 μM to low nanomolar (**Table 1** and **Supplementary Fig. 3**). As noted previously[14], there was little correlation between potency and SEA *E* value. Potency is ignored in weighting the ligand similarities in SEA; *E* values only indicate the likelihood that the compound will be active at a relevant concentration[13,16].

It is appropriate to consider the accuracy and the novelty of the predictions. Arguably, any method that uses a library of ligand-target sets such as ChEMBL and a metric of chemical similarity could have predicted targets for some of these molecules. For instance, pairwise compound similarity alone, or one of the more sophisticated methods now available[7,11,12,21,22], might well have suggested that compounds **6** and **7** (**Supplementary Table 3**) target the Kv1.2 potassium channel or that **8** targets vasopressin receptors. Any such approach must confront the problem of what level of chemical similarity usefully identifies likely targets; if a similarity cutoff is too permissive, too many targets will be captured, and if it is too stringent, many likely targets will be missed. For instance, the pairwise similarity of **2** for β-adrenergic receptor ligands and of **3** for α_2-adrenergic receptor ligands never rose above a T_C of 0.61 (Daylight fingerprints) or 0.43 (ECFP4 fingerprints), respectively. At these relatively low similarities, a simple two-dimensional similarity search of ChEMBL predicts 217 targets for **2** and 57 targets for **3**, many of which will be false positives. SEA attempts to address this problem by comparing observed similarities to those expected at random and by screening any bait molecule against an entire ligand-target set[13,16]. This at once increases the number of addressable targets, ranks the predictions by confidence level and eliminates many predictions that are based on a bait molecule's association with one or two ligands in a target set that might annotate hundreds of molecules. Because SEA is model free and uses all information represented in a fingerprint without weighting, it can interrogate any ligand-target set for similarity to any bait molecule without defining warheads or pharmacophores.

A key challenge, for us and for the field, is linking the targets for which *in vitro* affinity is measured to the *in vivo* phenotype that was originally observed. One way to investigate this is to test another molecule, known to be active on the target but structurally dissimilar to the active compound from the phenotypic screen, for its ability to phenocopy the original bait molecule in the animal or cell[23]. The voltage-gated potassium channel Kv1.2 is attractive for this purpose in that it is modulated by structurally diverse molecules. Psora-4 (**12**), for instance, has low- to mid-nanomolar activity against members of the Kv1 family, including Kv1.2, but is structurally orthogonal to compounds **6** and **7**. Consistent with the prediction that compounds **6** and **7** exert their excitatory phenotype by blocking this family of ion channels, the PMR profile induced by the chemically unrelated compound **12** closely resembles that induced by compounds **6** and **7** (**Fig. 2** and **Supplementary Fig. 4**).

A second pharmacological test for a target-based effect is to use the screening compound to compete against the function of a known ligand for the target of interest. We tested the ability of compound **1** to reverse the excitatory activity of the β-adrenergic agonist isoproterenol (**13**) in the PMR assay. Consistent with the prediction and *in vitro* observation of activity on the β-adrenergic receptors, compound **1** reversed isoproterenol's excitatory PMR as

Table 1 | Successful target predictions for compounds active in the zebrafish PMR screen

Compound	Predicted targets	*E* value	K_i (nM)
1	β_1-AR	2.99×10^{-18}	337
	β_2-AR	4.69×10^{-15}	728
	β_3-AR	4.55×10^{-19}	589
2	β_1-AR	—	965
	β_2-AR	2.98×10^{-5} [a]	550
	β_3-AR	2.98×10^{-5} [a]	3,128
3	D2R	3.54×10^{-33}	120
	D3R	8.16×10^{-41}	>10,000
	D4R	1.00×10^{-24}	5.0
	α_{1A}-AR	2.07×10^{-48}	113.6
	α_{1B}-AR	1.19×10^{-48}	29.7
	α_{1D}-AR	3.89×10^{-48}	25.4
	α_{2A}-AR	2.01×10^{-27}	164
	α_{2B}-AR	2.83×10^{-26}	96.4
	α_{2C}-AR	1.84×10^{-26}	80.8
	5-HT1A	3.83×10^{-67}	65.0
	5-HT1B	1.62×10^{-16}	9,422
	5-HT1D	7.55×10^{-17}	153
4	D2R	2.78×10^{-42}	4,869
	D3R	1.46×10^{-46}	946
	D4R	2.25×10^{-49}	1181
	5-HT1A	1.24×10^{-31}	2,439
	5-HT1B	1.61×10^{-12}	>10,000
	5-HT1D	2.32×10^{-16}	4,512
5 (ARC-239)	D2R	2.15×10^{-38}	207
	D3R	9.25×10^{-44}	164
	D4R	8.28×10^{-25}	530
6	Kv1.1	4.05×10^{-71}	n.d.
	Kv1.2	9.40×10^{-85}	1,930[b]
	Kv1.3	1.28×10^{-14}	n.d.
7	Kv1.1	9.69×10^{-56}	n.d.
	Kv1.2	1.18×10^{-66}	2,810[b]
	Kv1.3	1.74×10^{-10}	n.d.
8	OXTRA	1.60×10^{-26}	>10,000
	AVPR1A	8.32×10^{-10}	690
9 (NAN-190)	SERT	2.06×10^{-19}	6,200[c]
10	ABL2	1.62×10^{-54}	7,460[b,d]
	SRC	3.38×10^{-34}	3,820[b,d]
	LCK	2.51×10^{-30}	1,890[b,d]
11	p38α	2.68×10^{-10}	47%[e]

[a]Calculated with Daylight fingerprints. [b]IC_{50} (nM). [c]EC_{50} (nM). [d]Assuming competitive inhibition, and on the basis of the Cheng-Prusoff equation, K_i values would be about half the IC_{50} values for the kinase inhibitors. [e]Percentage inhibition at 10 μM. AR, adrenergic receptor. Chemical structures and PMR plots of these compounds are shown in **Supplementary Table 3**.

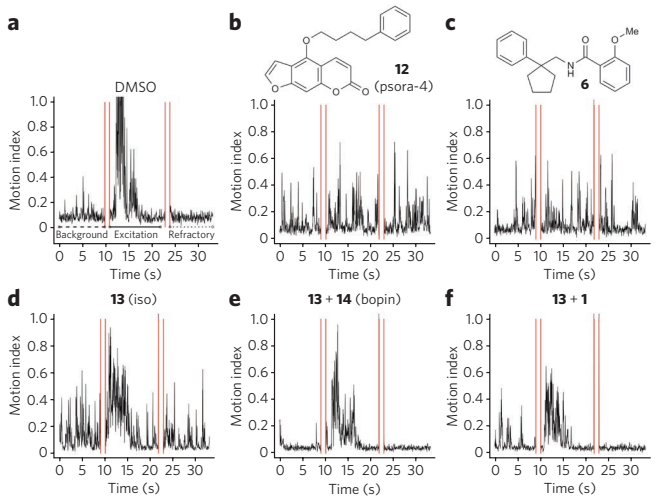

Figure 2 | Testing target relevance by phenocopy and functional competition. Compounds **6** and **7** phenocopy the known Kv inhibitor psora-4 (**12**). (**a**–**c**): The PMR phenotypes of animals treated with DMSO (**a**), compound **12** (**b**) or compound **6** (**c**). The same phenotype was obtained for compound **7** (detailed in **Supplementary Table 3**). Compound **1** suppresses β-adrenergic receptor agonist-induced motor excitation (**d**–**f**): PMR phenotypes of animals treated with: compound **13** (iso, isoproterenol) (**d**), compound **13** and the known β-adrenergic receptor antagonist compound **14** (bopin, bopindolol) (**e**), or compounds **13** and **1** (**f**).

effectively as the well-known β-adrenergic antagonist bopindolol (**Fig. 2** and **Supplementary Fig. 5**). In contrast, **1** did not reverse the excitatory phenotype of drugs acting on other receptor classes, such as the digitoxigenins, nor did sedative drugs from other classes reverse the excitatory activity of isoproterenol (**Supplementary Fig. 6**). These results are consistent with the zebrafish phenotype of compound **1** being mediated via β-adrenergic receptors.

Chemoinformatic target identification in phenotypic screens has important advantages: it is rapid, has a relatively high success rate and can address both high- and low-abundance targets. Admittedly, there are important caveats: the approach is limited to liganded targets, and even for these, almost half of the predictions that were tested failed, as in earlier studies[13,14]. Still, this success rate seems high enough to be pragmatic for experimental prioritization. Especially when a purely empirical screen is laborious, and when low-abundance targets will be missed, this approach and related methods[11,12,22,24] will usefully complement purely experimental approaches. Although finding that a molecule is active against a particular target does not establish the role of that target in the phenotype, it does provide a testable molecular hypothesis. Whereas the method is restricted to the targets for which ligand information is available, one can at least be sure that these are ligandable and that small-molecule probes already exist within the small part of chemical space that has been explored[25]. Within this set, the method is systematic and comprehensive enough to suggest testable targets for most of the ligands that are active in even a diverse ligand library, as was tested here (**Table 1**). This method may thus find broad application to target identification in phenotypic screens, and we have developed a website from which the method may be accessed by the community (http://sea.bkslab.org/).

Received 25 April 2011; accepted 19 September 2011; published online 18 December 2011

References

1. Harding, M.W., Galat, A., Uehling, D.E. & Schreiber, S.L. *Nature* **341**, 758–760 (1989).
2. Leung, D., Hardouin, C., Boger, D.L. & Cravatt, B.F. *Nat. Biotechnol.* **21**, 687–691 (2003).
3. Ong, S.E. *et al. Proc. Natl. Acad. Sci. USA* **106**, 4617–4622 (2009).
4. Mayer, T.U. *et al. Science* **286**, 971–974 (1999).
5. Haggarty, S.J. *et al. Chem. Biol.* **7**, 275–286 (2000).
6. Schreiber, S.L. *Nat. Chem. Biol.* **1**, 64–66 (2005).
7. Seiler, K.P. *et al. Nucleic Acids Res.* **36**, D351–D359 (2008).
8. Palchaudhuri, R. & Hergenrother, P.J. *ACS Chem. Biol.* **6**, 21–33 (2011).
9. Lamb, J. *et al. Science* **313**, 1929–1935 (2006).
10. Keiser, M.J., Irwin, J.J. & Shoichet, B.K. *Biochemistry* **49**, 10267–10276 (2010).
11. Paolini, G.V., Shapland, R.H., van Hoorn, W.P., Mason, J.S. & Hopkins, A.L. *Nat. Biotechnol.* **24**, 805–815 (2006).
12. Yildirim, M.A., Goh, K.I., Cusick, M.E., Barabasi, A.L. & Vidal, M. *Nat. Biotechnol.* **25**, 1119–1126 (2007).
13. Keiser, M.J. *et al. Nat. Biotechnol.* **25**, 197–206 (2007).
14. Keiser, M.J. *et al. Nature* **462**, 175–181 (2009).
15. Kokel, D. *et al. Nat. Chem. Biol.* **6**, 231–237 (2010).
16. Hert, J., Keiser, M.J., Irwin, J.J., Oprea, T.I. & Shoichet, B.K. *J. Chem. Inf. Model.* **48**, 755–765 (2008).
17. Rogers, D. & Hahn, M. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
18. James, C., Weininger, D. & Delany, J. *Daylight Theory Manual* (Daylight Chemical Information Systems Inc., 2011).
19. Willett, P., Barnard, J.M. & Downs, G.M. *J. Chem. Inf. Comput. Sci.* **38**, 983–996 (1998).
20. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. *J. Mol. Biol.* **215**, 403–410 (1990).
21. Young, D.W. *et al. Nat. Chem. Biol.* **4**, 59–68 (2008).
22. Mestres, J., Gregori-Puigjane, E., Valverde, S. & Sole, R.V. *Mol. Biosyst.* **5**, 1051–1057 (2009).
23. Alderton, W., Davenport, R. & Fish, P.V. *Drugs Future* **35**, 517–521 (2010).
24. Kinnings, S.L. *et al. PLOS Comput. Biol.* **5**, e1000423 (2009).
25. Hert, J., Irwin, J.J., Laggner, C., Keiser, M.J. & Shoichet, B.K. *Nat. Chem. Biol.* **5**, 479–483 (2009).

Acknowledgments

We thank S. Morris for help with Cytoscape. This work was supported by US National Institutes of Health grants GM71896 (to J.J.I. and B.K.S.), AG02132 (to S. Prusiner and B.K.S.), MH085205 and MH086867 (to R.P.), MH091449 (to D.K.), R01 MH093603 and R01 NS49272 (to D.L.M.); a Rogers Family Foundation award (to M.J.K.); the National Institutes of Mental Health Psychoactive Drug Screening Program, grant U19MH82441; the Michael Hooker Chair (to B.L.R.); and fellowships from the Max Kade Foundation (to C.L.) and the European Molecular Biology Organization (to A.T.).

Author contributions

The strategy was devised by B.K.S. and R.T.P., the PMR assay by D.K., and target predictions and other calculations by C.L., with assistance and editing by J.J.I., M.J.K., H.L. and B.K.S. Electrophysiology was designed by D.L.M. and implemented by A.T., and GPCR and kinase experiments were designed and implemented by B.L.R. and V.S., who also advised on target-phenotype associations. Zebrafish pharmacology was conducted by D.K. with assistance by C.Y.J.C.

Competing financial interests

The authors declare no competing financial interests.

Additional information

Supplementary information and chemical compound information is available online at http://www.nature.com/naturechemicalbiology/. Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Correspondence and requests for materials should be addressed to B.L.R., R.T.P. or B.K.S.