Distinct biological network properties between the targets of natural products and disease genes

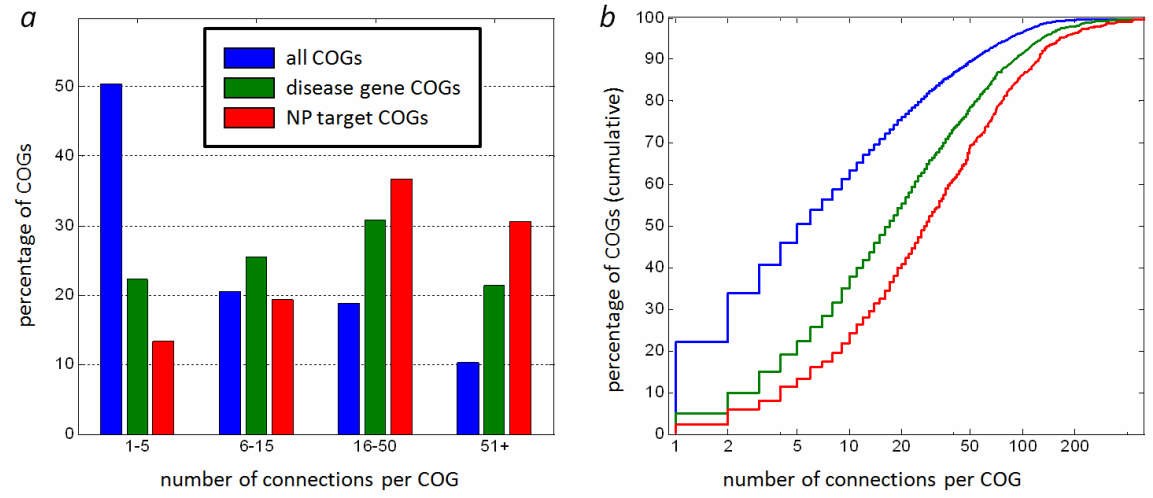
Vlado Dančík[‡], Kathleen Petri Seiler, Damian W. Young, Stuart L. Schreiber[*], Paul A. Clemons[*]

Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, MA 02143

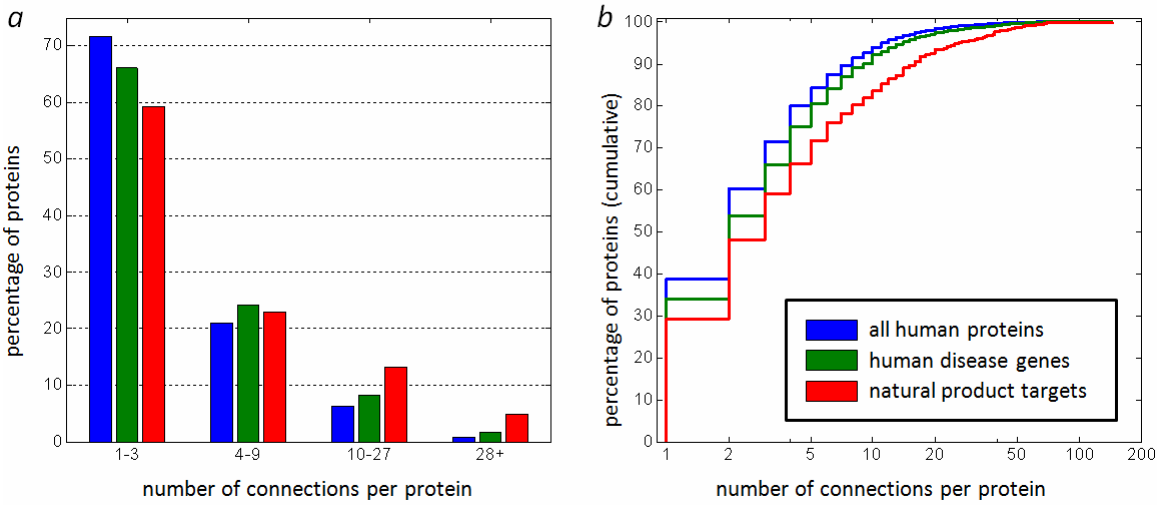
Supporting Information. This document contains supplementary analyses referenced in the main narrative, including five Figures (**Supplementary Figures S1-S5**). This document also contains the details of Statistical Methods used in this study.

Supplementary Analyses

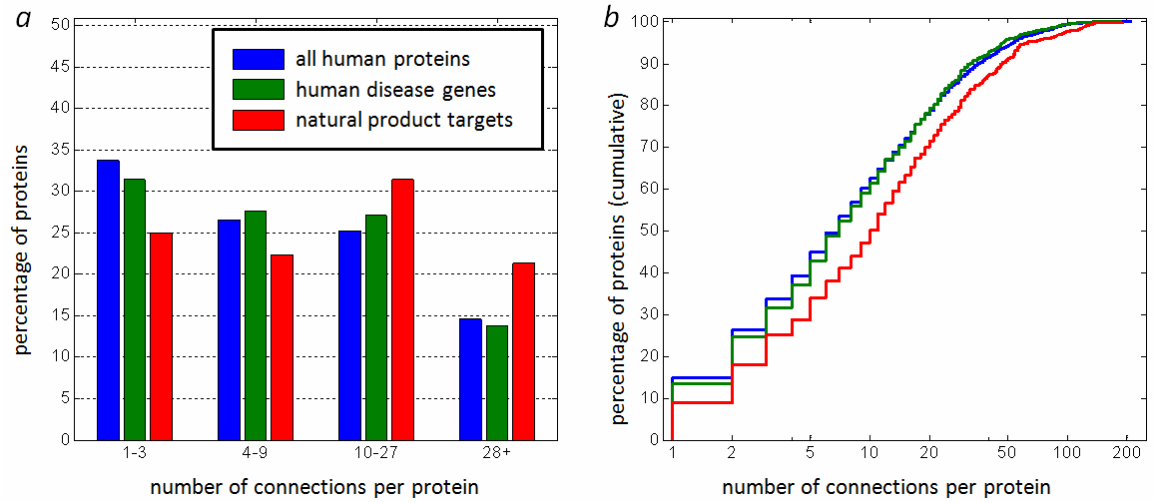
As control experiments described in the text, we analyzed STRING protein connections in three additional ways. First, we assessed the distribution of protein connectivities using natural product target proteins belonging to clusters of orthologous groups (COGs, http://www.ncbi.nlm.nih.gov/COG/) as defined by the STRING database[1,2] as network nodes. As with the primary analysis in the main narrative, we considered proteins with at least one connection among all STRING protein COGs, the natural product targets mapped to COGs, and heritable disease gene COGs. Second, we established protein connections using experimental evidence only. For both of these comparisons (Figures S1 and S2, respectively), the distributions and relative relationships of these groups to one another mimicked the outcomes using individual proteins and all types of STRING evidence (Figure 1 in the main narrative). We also considered STRING connections obtained by mining manually curated pathway databases. In this case, we observe closer correspondence between disease genes and all proteins, with natural products remaining more highly connected than either (Figure S3). To control for the possibility that GVKBio natural product targets are skewed by targets only weakly inhibited, we re-analyzed our results using only those interactions in GVKBio reported to be in the low nanomolar range or below ($< 10^{-7.5}$ *M*); these results look very similar to those using all GVKBio records (Figure S4). Finally, since a large fraction of drug targets are G protein-coupled receptors[3], we looked at the overall connectivities of GPCRs in STRING by identifying the subset of STRING annotated as GPCRs using GPCRDB (http://www.gpcr.org/7tm/). Interestingly, this analysis shows that GPCRs are much less connected than all proteins in STRING (Figure S5).



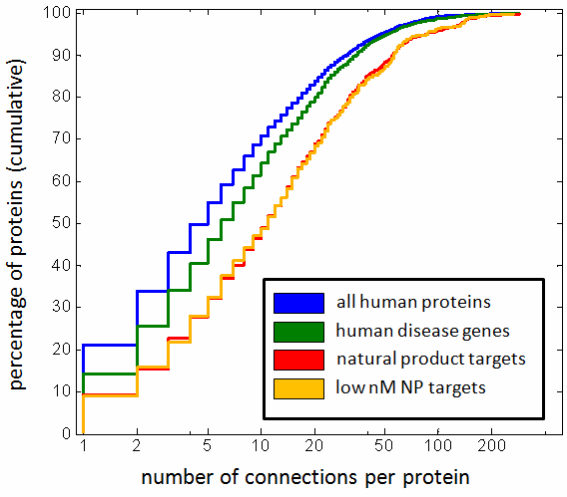
Supplementary Figure S1. *Network connectivity of COGs.* Since many natural products target non-human proteins we also performed analysis with the COG (Clusters of Orthologous Groups of proteins, http://www.ncbi.nlm.nih.gov/COG/) network available from STRING. (a) Connectivity summary of different COGs: all COGs in STRING database (blue: n = 13,091; median = 5; mean = 18.4), COGs containing disease-associated proteins (green: n = 2,597; median = 17; mean = 35.7), COGs containing natural product targets (red: n = 819; median = 28; mean = 51.1). (b) Cumulative connectivity distributions illustrating differences.



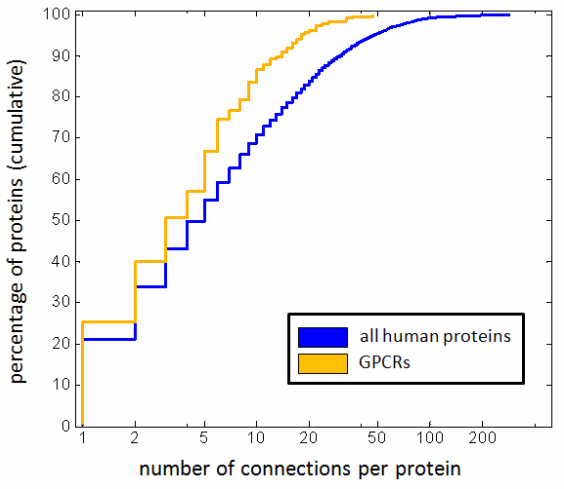
Supplementary Figure S2. *Network connectivity using experimental evidence only to make STRING connections.* (a) Connectivity summary of different target groups: all proteins in STRING database (blue: n = 4,612; median = 2; mean = 3.6), disease-associated proteins (green: n = 1,469; median = 2; mean = 4.3), natural product targets (red: n = 499; median = 3; mean = 6.6). (b) Cumulative connectivity distributions illustrating differences.



Supplementary Figure S3. *Network connectivity using curated pathway database evidence only to make STRING connections.* (a) Connectivity summary of different target groups: all proteins in STRING database (blue: n = 4,949; median = 7; mean = 13.9), disease-associated proteins (green: n = 1,621; median = 7; mean = 13.5), natural product targets (red: n = 780; median = 10; mean = 18.7). (b) Cumulative connectivity distributions illustrating differences.



Supplementary Figure S4. *Selection of only high-potency natural product-target interactions from GVKBio database.* Cumulative connectivity distribution comparison of target list from interactions with $EC_{50} < 10^{-7.5}$ *M* in GVKBio database (gold: n = 483; median = 11; mean = 22.5) with all proteins in STRING (blue; same as Figure 1b), natural product targets (red; same as Figure 1b), and human disease genes (green; same as Figure 1b).



Supplementary Figure S5. *Analysis of connectivity of G protein-coupled receptors (GPCRs) in STRING database.* Cumulative connectivity distribution comparison of proteins in STRING identified as GPCRs by GPCRDB (http://www.gpcr.org/7tm/; gold: n = 233; median = 3; mean = 5.7) with all proteins in STRING (blue; same as Figure 1b).

Statistical Methods

We used statistical analysis to assess the significance of the differences between the connectivity distributions of natural product targets, disease genes, and all proteins in the STRING network. The connectivity distributions are not normally distributed as is illustrated by the differences between the mean and median values (Table 1) and are typically modeled by a power-law distribution[4]. Due to lack of normality, we used a nonparametric Kolmogorov-Smirnov goodness-of-fit test as implemented in a MATLAB statistical toolbox. Most comparisons were significant at significance level $\alpha = 0.05$ (Table 2). No adjustments for multiple testing were applied.

Network	Source	number	mean	median	std dev
STRING	STRING proteins	8799	11.7	5	19.6
STRING	Disease genes	2681	14.0	6	22.3
STRING	NP targets	946	22.5	11	33.4
STRING	Manually selected	38	48.2	32.5	44.5
STRING	ChEMBL*	729	17.4	8	24.8
STRING	DrugBank**	731	14.9	7	21.6
STRING	Low nM NP targets	483	22.5	11	32.7
STRING	GPCRs	233	5.7	3	6.6
COG	STRING	13091	18.4	5	34.1
COG	Disease genes	2597	35.7	17	53.5
COG	NP targets	819	51.1	28	70.2
Experimental	STRING	4612	3.6	2	5.5
Experimental	Disease genes	1469	4.3	2	7.3
Experimental	NP targets	499	6.6	3	11.6
Pathways	STRING	4949	13.9	7	19.1
Pathways	Disease genes	1621	13.5	7	17.7
Pathways	NP targets	780	18.7	10	24.4

Table 1. The number of proteins, mean, median, and standard deviation of connectivities. *ChEMBL – targets of bioactive molecules from ChEMBL excluding NP targets. **DrugBank – targets of approved molecules from DrugBank excluding NP targets.

Figure	Comparison		Network	Two-sided p-value	One-sided p-value
F1	STRING proteins	Disease genes	STRING	2.78E-15	1.39E-15
F1	STRING proteins	NP targets	STRING	1.91E-39	9.53E-40
F1	Disease genes	NP targets	STRING	2.57E-15	1.28E-15
F2	NP targets	Manually selected	STRING	7.85E-06	3.93E-06
F3	Disease genes	ChEMBL*	STRING	0.00016816	8.41E-05
F3	ChEMBL*	NP targets	STRING	0.014087	0.0070437
F4	*Disease genes*	*DrugBank***	*STRING*	*0.18604*	*0.093097*
F4	DrugBank**	NP targets	STRING	1.92E-05	9.58E-06
S1	STRING proteins	Disease genes	COG	1.42E-149	7.11E-150
S1	STRING proteins	NP targets	COG	8.99E-106	4.50E-106
S1	Disease genes	NP targets	COG	3.54E-13	1.77E-13
S2	STRING proteins	Disease genes	Experimental	0.00025319	0.0001266
S2	STRING proteins	NP targets	Experimental	5.07E-08	2.53E-08
S2	Disease genes	NP targets	Experimental	0.0051732	0.0025866
S3	*STRING proteins*	*Disease genes*	*Pathways*	*0.57661*	*0.29596*
S3	STRING proteins	NP targets	Pathways	1.88E-10	9.39E-11
S3	Disease genes	NP targets	Pathways	4.64E-07	2.32E-07
S4	STRING proteins	Low nM NP targets	STRING	9.80E-21	4.90E-21
S4	Disease genes	Low nM NP targets	STRING	2.93E-09	1.47E-09
S4	*NP targets*	*Low nM NP targets*	*STRING*	*1*	*0.89336*
S5	GPCRs	STRING proteins	STRING	1.88E-05	9.38E-06

Table 2. P-values for Kolmogorov-Smirnov tests for two samples ([5], Test 13). Three comparisons that fail to achieve statistical significance at level $\alpha = 0.05$ are italicized. *ChEMBL – targets of bioactive molecules from ChEMBL excluding NP targets. **DrugBank – targets of approved molecules from DrugBank excluding NP targets.

Manually curated compounds

As described in the text, we hand-curated 76 natural products ([6-10] and references therein) and mapped their targets to STRING (compounds highlighted in yellow are depicted in Figure 2b).

CompoundName	ProteinTarget(s)	Gene	Ensembl	Connections
verrucarin A	MAP kinase inhibitor	MAPK1	ENSP00000215832	182
butyrolactone-I	Cdk1/CycB	CDC2	ENSP00000362917	144
hymenialdisine	Cdk1/CycB, CDK5/p25, GSK3B, Mek1	CDC2	ENSP00000362917	144
oleanolic acid	NF-KB, STAT3, STAT5 inhibition	STAT3	ENSP00000264657	144
aparatoxin A	JAK/STAT	JAK1	ENSP00000294423	127
oleanolic acid	NF-KB, STAT3, STAT5 inhibition	NFKB1	ENSP00000226574	111
bryostatin	protein kinase C inhibtion	PRKCA	ENSP00000284384	86
genistein	BCL-2	PRKCA	ENSP00000284384	86
ingenol	protein kinase C activation	PRKCA	ENSP00000284384	86
rebeccamycin	topisomerase I and II	PRKCA	ENSP00000284384	86
staurosporine	FLT3 inhibition	PRKCA	ENSP00000284384	86
variolin B	CDK inhibitor	CDKN1A	ENSP00000362816	82
cryptophycins	80S ribosome and 60S ribosomal subunit	RPL7	ENSP00000339795	69
butyrolactone-I	Cdk1/CycB	CCNB1	ENSP00000370233	66
hymenialdisine	Cdk1/CycB, CDK5/p25, GSK3B, Mek1	CCNB1	ENSP00000370233	66
hymenialdisine	Cdk1/CycB, CDK5/p25, GSK3B, Mek1	MAP2K1	ENSP00000302486	63
oleanolic acid	NF-KB, STAT3, STAT5 inhibition	STAT5B	ENSP00000293328	62
FK506	FKBP12/calcineurin	MTOR	ENSP00000354587	61
rapamycin		MTOR	ENSP00000354587	61
didemnin B		EIF4A1	ENSP00000369881	60
plitidepsin	VEGF and VEGF1 inhibitor	EIF4A1	ENSP00000369881	60
epoxomycin	proteasome	PSMD1	ENSP00000309474	56
fellutamide B	proteasome	PSMD1	ENSP00000309474	56
lactacystin	proteasome	PSMD1	ENSP00000309474	56
salinosporamide A	proteasome inhibitor	PSMD1	ENSP00000309474	56
hymenialdisine	Cdk1/CycB, CDK5/p25, GSK3B, Mek1	GSK3B	ENSP00000324806	53
apicidin	HDAC inhibition	HDAC3	ENSP00000302967	52
butyric acid	HDAC inhibition	HDAC3	ENSP00000302967	52
FK228	HDAC inhibition	HDAC3	ENSP00000302967	52
psammaplin A	HDAC inhibition	HDAC3	ENSP00000302967	52
trapoxin	HDAC inhibition	HDAC3	ENSP00000302967	52
trichostatin	HDAC inhibition	HDAC3	ENSP00000302967	52
dictyodendrin A	telomerase	TERT	ENSP00000309572	50
dehydroaltenusin	mammalian DNA polymerase alpha	POLA1	ENSP00000368358	35
caliculin		PPP3CA	ENSP00000320580	33
cyclosporin	cyclophilin/calcineurin	PPP3CA	ENSP00000320580	33
FK506	FKBP12/calcineurin	PPP3CA	ENSP00000320580	33
fostriecin	PP2A and PP4	PPP2CA	ENSP00000231504	32
okadaic acid	PP1 and PP2A	PPP2CA	ENSP00000231504	32

10-deacetyl baccatin III	tubulin stabilization	TUBB1	ENSP00000217133	29
colchicine	tubulin binder	TUBB1	ENSP00000217133	29
combretastatin A-4	tubulin binding	TUBB1	ENSP00000217133	29
curacin A	tubulin	TUBB1	ENSP00000217133	29
cytochalasin A	tubulin assembly inhibition	TUBB1	ENSP00000217133	29
discodermolide	tubulin assembly inhibition	TUBB1	ENSP00000217133	29
dolastatin	tubulin assembly inhibition	TUBB1	ENSP00000217133	29
epothilone A	alpha-beta tubulin	TUBB1	ENSP00000217133	29
halichondrin B	tubulin assembly inhibition	TUBB1	ENSP00000217133	29
hemiasterlin	tubulin assembly inhibition	TUBB1	ENSP00000217133	29
noscapine	tubulin binding	TUBB1	ENSP00000217133	29
paclitaxel	tubulin stabilization	TUBB1	ENSP00000217133	29
spongistatin	tubulin binder	TUBB1	ENSP00000217133	29
taxol	alpha-beta tubulin	TUBB1	ENSP00000217133	29
vinblastine	tubulin binding	TUBB1	ENSP00000217133	29
vincristine		TUBB1	ENSP00000217133	29
hymenialdisine	Cdk1/CycB, CDK5/p25, GSK3B, Mek1	CDK5	ENSP00000297518	28
spisulosine	GTP-binding protein RHO	RHO	ENSP00000296271	24
ZM447439	aurora kinase	AURKA	ENSP00000321591	24
jasplakinolide	actin	ACTB	ENSP00000349960	22
latrunculin		ACTB	ENSP00000349960	22
phallodidin	actin	ACTB	ENSP00000349960	22
cyclopamine	SMO	SMO	ENSP00000249373	21
avrainvillamide	nucleophosmin	NPM1	ENSP00000296930	19
morphine	mu opioid receptor	OPRM1	ENSP00000229768	18
camptothecin	DNA topoisomerase	TOP1	ENSP00000354522	17
epipodophyllotoxin	topisomerase I and II	TOP1	ENSP00000354522	17
rebeccamycin	topisomerase I and II	TOP1	ENSP00000354522	17
tryprostatin	Map2	MAP2	ENSP00000353508	16
pladieolide B	SFB3 subunit 3	SF3B3	ENSP00000305790	15
nicotine	nicotinic acetylcholine receptor	CHRNA7	ENSP00000303727	11
daidzein	NADH oxidase(tNOX inhibition)	NOX1	ENSP00000362057	8
cyclosporin	cyclophilin/calcineurin	PPIA	ENSP00000348240	3
myriocin		SPTLC1	ENSP00000262554	3
fumagillin	METAP2 inhibition	METAP2	ENSP00000325312	2
cryptophycins	80S ribosome and 60S ribosomal subunit	EIF2A	ENSP00000273435	1
penicillin	beta-lactamase	LACTB	ENSP00000261893	1

References

(1) Jensen, L. J.; Kuhn, M.; Stark, M.; Chaffron, S.; Creevey, C.; Muller, J.; Doerks, T.; Julien, P.; Roth, A.; Simonovic, M.; Bork, P.; von Mering, C. *Nucleic Acids Res* **2009**, *37*, D412-6.

(2) von Mering, C.; Jensen, L. J.; Snel, B.; Hooper, S. D.; Krupp, M.; Foglierini, M.; Jouffre, N.; Huynen, M. A.; Bork, P. *Nucleic Acids Res* **2005**, *33*, D433-7.

(3) Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. *Nat Rev Drug Discov* **2006**, *5*, 993-6.

(4) Barabasi, A. L.; Albert, R. *Science* **1999**, *286*, 509-12.

(5) Sheshkin, D. J. *Handbook of Parametric and Nonparametric Statistical Procedures*; 2 ed.; Chapman & Hall/CRC, 2004.

(6) Molinski, T. F.; Dalisay, D. S.; Lievens, S. L.; Saludes, J. P. *Nat Rev Drug Discov* **2009**, *8*, 69-85.

(7) Feldman, I.; Rzhetsky, A.; Vitkup, D. *Proc Natl Acad Sci U S A* **2008**, *105*, 4323-8.

(8) Newman, D. J.; Cragg, G. M. *Curr Drug Targets* **2006**, *7*, 279-304.

(9) Nagle, A.; Hur, W.; Gray, N. S. *Curr Drug Targets* **2006**, *7*, 305-26.

(10) Newman, D. J.; Cragg, G. M.; Holbeck, S.; Sausville, E. A. *Curr Cancer Drug Targets* **2002**, *2*, 279-308.