STITCH 4: integration of protein–chemical interactions with user data

Kuhn, Michael; Szklarczyk, Damian; Pletscher-Frankild, Sune; Blicher, Thomas H; von Mering, Christian; Jensen, Lars J; Bork, Peer

Abstract: STITCH is a database of protein-chemical interactions that integrates many sources of experimental and manually curated evidence with text-mining information and interaction predictions. Available at http://stitch.embl.de, the resulting interaction network includes 390 000 chemicals and 3.6 million proteins from 1133 organisms. Compared with the previous version, the number of high-confidence protein-chemical interactions in human has increased by 45%, to 367 000. In this version, we added features for users to upload their own data to STITCH in the form of internal identifiers, chemical structures or quantitative data. For example, a user can now upload a spreadsheet with screening hits to easily check which interactions are already known. To increase the coverage of STITCH, we expanded the text mining to include full-text articles and added a prediction method based on chemical structures. We further changed our scheme for transferring interactions between species to rely on orthology rather than protein similarity. This improves the performance within protein families, where scores are now transferred only to orthologous proteins, but not to paralogous proteins. STITCH can be accessed with a web-interface, an API and downloadable files

DOI: https://doi.org/10.1093/nar/gkt1207

Published online 28 November 2013 *Nucleic Acids Research, 2014, Vol. 42, Database issue* ***D401–D407***
doi:10.1093/nar/gkt1207

STITCH 4: integration of protein–chemical interactions with user data

Michael Kuhn[1,*], Damian Szklarczyk[2], Sune Pletscher-Frankild[3], Thomas H. Blicher[3], Christian von Mering[2], Lars J. Jensen[3,*] and Peer Bork[4,5,*]

[1]Biotechnology Center, TU Dresden, 01062 Dresden, Germany, [2]Institute of Molecular Life Sciences, University of Zurich and Swiss Institute of Bioinformatics, Winterthurerstrasse 190, 8057 Zurich, Switzerland, [3]Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, 2200 Copenhagen N, Denmark, [4]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany and [5]Max-Delbrück-Centre for Molecular Medicine, Robert-Rössle-Strasse 10, 13092 Berlin, Germany

Received September 30, 2013; Revised November 1, 2013; Accepted November 4, 2013

ABSTRACT

STITCH is a database of protein–chemical interactions that integrates many sources of experimental and manually curated evidence with text-mining information and interaction predictions. Available at http://stitch.embl.de, the resulting interaction network includes 390 000 chemicals and 3.6 million proteins from 1133 organisms. Compared with the previous version, the number of high-confidence protein–chemical interactions in human has increased by 45%, to 367 000. In this version, we added features for users to upload their own data to STITCH in the form of internal identifiers, chemical structures or quantitative data. For example, a user can now upload a spreadsheet with screening hits to easily check which interactions are already known. To increase the coverage of STITCH, we expanded the text mining to include full-text articles and added a prediction method based on chemical structures. We further changed our scheme for transferring interactions between species to rely on orthology rather than protein similarity. This improves the performance within protein families, where scores are now transferred only to orthologous proteins, but not to paralogous proteins. STITCH can be accessed with a web-interface, an API and downloadable files.

INTRODUCTION

Protein–chemical interactions are essential for any biological system; for example, they drive the metabolism of the cell or initiate many signaling cascades and most pharmaceutical interventions. A large collection of such interactions can, therefore, be used to study a variety of cellular functions and the impact of drug treatment on the cell. For such research, it is important to have, as complete as possible, data on protein–chemical interactions. By treating proteins and chemicals as nodes of a graph, which are linked by edges if they have been found to interact (1), we can adopt a network view that enables us to integrate many different sources. The concept of STITCH ('search tool for interacting chemicals') was from the beginning to combine sources of protein–chemical interactions from experimental databases, pathway databases, drug–target databases, text mining and drug–target predictions into a unified network (2–4). This network abstracts the complexity of the underlying data sources, making large-scale studies possible. At the same time, links to the original sources are retained, making it possible to trace the provenance of the data. The underlying STITCH database can be accessed in multiple ways: via an intuitive web interface, via download files (for large-scale analysis) and via an API (enabling automated access on a small to medium scale). Here, we present recent improvements to the database and user interface of STITCH. Already in the previous versions, it has been possible to query STITCH using protein or chemical names, InChIKeys and SMILES strings. New in this version is the possibility to upload spreadsheets with chemical descriptors and experimental data that can be directly added to the network, as described later in text. We also for the first time use the evidence transfer algorithm described for the STRING 9.1 database (5) to improve the performance for protein families.

Compared with STITCH 3, we use the same underlying set of proteins, containing 1133 species. We updated the

*To whom correspondence should be addressed. Tel: +49 6221 387 8526; Fax: +49 6221 387 8517; Email: bork@embl.de
Correspondence may also be addressed to Michael Kuhn. Tel: +49 351 463 40063; Fax: +49 351 463 40061; Email: michael.kuhn@biotec.tu-dresden.de
Correspondence may also be addressed to Lars J. Jensen. Tel: +45 353 25025; Fax: +45 353 25001; Email: lars.juhl.jensen@cpr.ku.dk

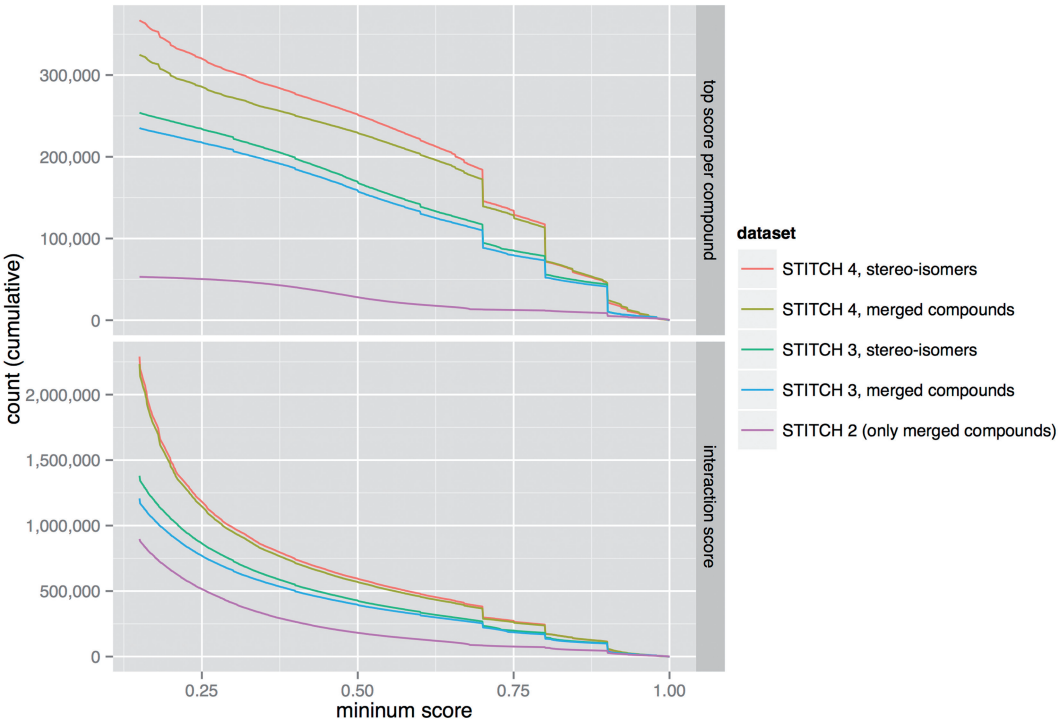


Figure 1. Cumulative distribution of scores. For each confidence score threshold, the plot shows the number of chemicals (top) and protein–chemical interactions (bottom) that have at least this confidence score in the human protein–chemical network. For example, there are 172 000 chemicals with a high-confidence interaction (score at least 0.7). As there are many interactions with low confidence scores, we use a minimum score threshold of 0.15. Steps in the data correspond to large numbers of compounds that have the same maximum score in manually curated databases or the ChEMBL database (with different confidence levels).

set of chemicals (6), and find interactions with 390 000 distinct chemicals. In human, high-confidence interactions for 172 000 compounds are available in STITCH 4 (Figure 1), compared with 110 000 in STITCH 3 (4). In total, the human protein–chemical interaction network contains 2.2 million interactions (Figure 1). Applying different confidence thresholds, 570 000 interactions are of medium confidence (score cutoff 0.5) and 367 000 interactions are of high confidence (cutoff 0.7).

SOURCES OF INTERACTIONS

Protein–chemical interactions are presented in four different channels: experiments, databases, text mining and predicted interactions. We import the following sources of experimental information: ChEMBL [interactions with reported K_i or IC_{50} (7)], PDSP K_i Database (8), PDB (9) and—new to STITCH—data from two large-scale studies on kinase–ligand interactions (10,11). From the latter studies, we extracted 74 291 interactions between 229 compounds and 414 human kinases. We converted the reported residual kinase activities (10) and kinase affinities (11) to probabilistic scores, which gave rise to 14 187, 9431 and 5977 interactions of at least low, medium and high confidence, respectively. The second channel is made up of manually curated drug–target databases: DrugBank (12), GLIDA (13), Matador (14), TTD (15) and CTD (16); and pathway databases: KEGG (17), NCI/Nature Pathway Interaction Database (18), Reactome (19) and BioCyc (20).

PREDICTION OF INTERACTIONS

STITCH contains verified interactions (from the sources listed earlier in text) and predicted interactions, based on text mining and other prediction methods. In the text-mining channels, interactions were extracted from the literature using both co-occurrence text mining and Natural Language Processing (21,22). For the first time for STITCH, we not only use data from MEDLINE abstracts and OMIM (23) but also from full-text articles freely available from PubMed Central or publishers' Web sites.

In previous versions, we have used medical subject headings (MeSH) terms in text mining and when importing external databases. These terms allowed us to expand concepts like 'alpha adrenergic receptors' to individual proteins. We used to map MeSH terms to proteins using a combination of automatic and manual approaches, which led to errors in some cases. Furthermore, the mapping was only valid for human proteins. We have, therefore, started to use terms from the Gene Ontology [GO terms, (24)] to define groups of proteins. We excluded GO annotations based on mutant phenotypes (IMP) and electronic annotations (IEA). We then checked the coverage of GO annotations for all species in STITCH. We only mapped GO terms to proteins for species where at least 10% of the proteins have been annotated, namely *Drosophila melanogaster*, *Escherichia coli*, *Homo sapiens*, *Mus musculus*, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*.

As the coverage of synonyms is lower than for MeSH terms, we manually added additional synonyms to GO

terms to increase the text-mining sensitivity. As one GO term corresponds to multiple proteins, the resulting confidence score for the individual protein–chemical interactions should be down-weighted compared with interactions that are directly associated with a single protein. We, therefore, determined a correction factor through benchmarking (as a function of the number of member proteins in the GO term). For each channel, we looked at the GO terms that are interacting with chemicals. We then checked if the member proteins that are part of the GO terms are in turn interacting with chemicals. For each of these chemicals, we determined the fraction of member proteins that are interacting. For example, if a drug was known to bind two of the three α2-adrenergic receptors, it was added as a data point (x = 3, y = 2/3) to the benchmark data. The data points were then fitted for each channel by the following function:

$$f(x) = (x - a)^{-e^{\frac{b}{x}}}$$

For larger groups, the function approaches x^{-1} (i.e. interacting with one protein is not predictive for the other proteins).

In this version of STITCH, we introduced a fourth channel, namely predicted protein–chemical interactions based on chemical structure. Countless articles on the prediction of drug–target interactions have been published in the last years [e.g. (25–27), reviewed in (28)]. In many cases, however, the actual predictions are not available. We, therefore, implemented a relatively simple and transparent prediction scheme based on Random Forests (29,30): for each target for which >100 binding partners are known from the ChEMBL database, we attempted to make a prediction. To avoid biases, we first excluded highly similar chemicals, enforcing a maximum Tanimoto similarity of 0.9 (using Algorithm 2 described by Hobohm) (31) using 2D chemical fingerprints calculated with the chemistry development kit (32,33). We then added ten times as many random chemicals as non-binders to the training set and used the fingerprints as predictors for all compounds. Using 10-fold cross-validation, we assessed how predictive the model is (by calculating the Pearson correlation coefficient between the training data and the cross-validation results). We used the correlation as a correction factor to decrease the confidence score of the predicted interactions, which were predicted for all compounds occurring in the ChEMBL database. We repeated this procedure three times for each compound and used the median predicted score, to decrease the effect of the random negative set. As interactions were predicted from the experimental channel, the predictions and experimental channels are not independent of each other. To compute the combined score (which is shown on the network), we therefore took the highest of either score, instead of combining the scores in a Bayesian fashion as it is done for the other channels. In total, predictions were made for 767 proteins across 15 species. The median correlation between the training data and the cross-validation prediction was 0.90.

Links between compounds were also extracted from the aforementioned sources, if possible. (e.g. chemical reactions from pathway databases or co-mentioned chemicals from text mining.) We also predicted shared mechanisms of action from MeSH pharmacological actions, the Connectivity Map using the DIPS method (34), which tests for similar changes in gene expression on compound treatment, and from screening data from the Developmental Therapeutics Program NCI/NIH (35). The latter screening data replaces our previous analysis of the NCI60 panel. We considered only the 70 of 115 cell lines against which >10 000 compounds have been screened and centered the negative logarithm of GI50 values with respect to both compounds and cell lines. For the 47 692 compounds in the data set, we calculated all-against-all covariance across cell lines and converted these to probabilistic scores. This resulted in 114 072, 24 889 and 6890 pairs of compounds of at least low, medium and high confidence, respectively.

To account for the fact that many interactions are determined in model species, we transfer interactions between species. Previously, the sequence similarity between two proteins was used to determine the confidence in the transferred score. This had the disadvantage that when transferring evidence from a selective binder (e.g. inhibiting only one subtype of a receptor), all subtypes of the receptor in the target species would receive a similar score. In the new scheme, only the orthologous protein receives the evidence from the specific compound.

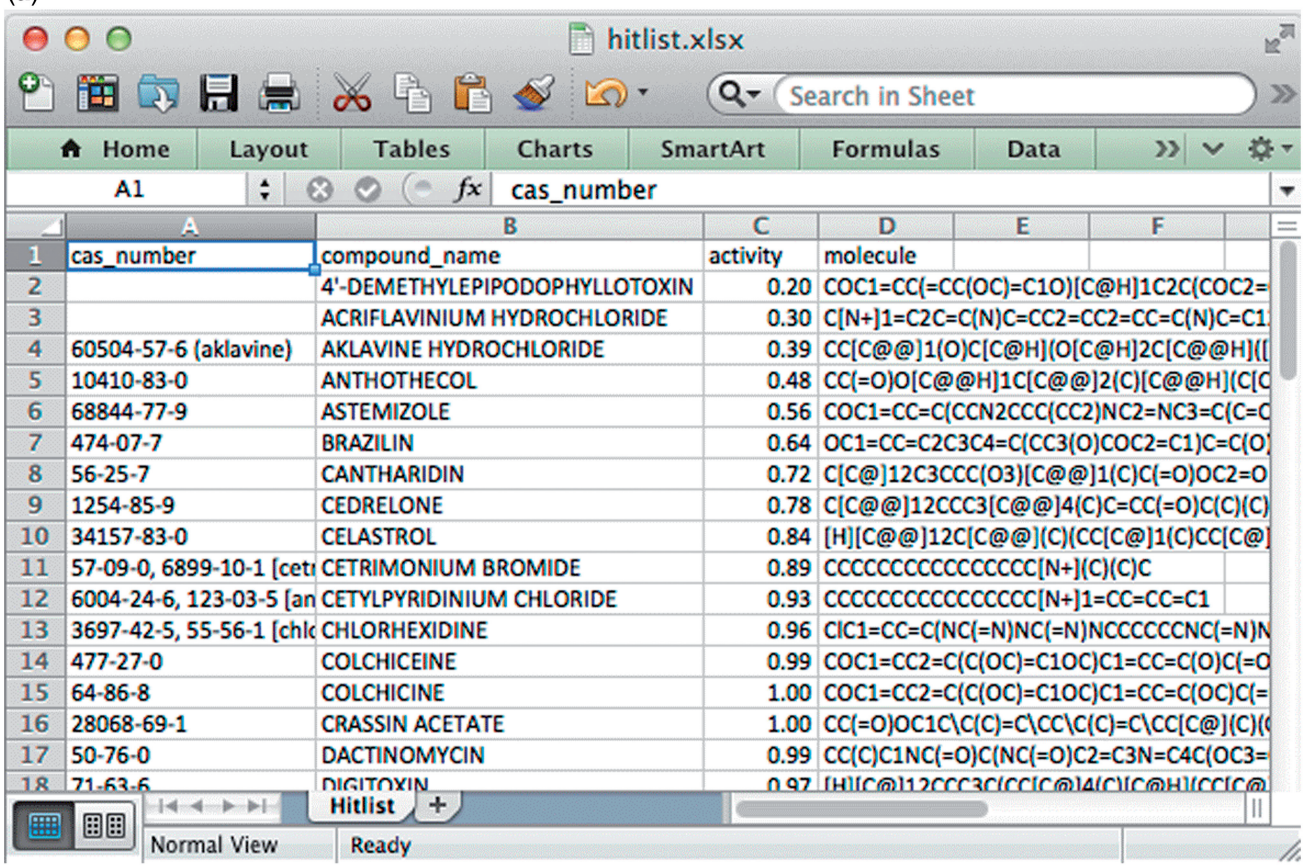
INTEGRATION WITH USER DATA

Users can now upload a spreadsheet (e.g. in Microsoft Excel format) with experimental data to STITCH using the 'batch import' functionality (Figure 2). For each compound, the spreadsheet may contain: the name of the compound, the chemical structure (as SMILES string, InChI or InChIKey), an internal identifier and a readout value. STITCH uses the name and chemical structure to find the compound in the STITCH database. The name provided by the user can then be shown in the interaction network, and the downloadable files contain both the name and the user's internal identifier (if provided). The readout value may be a numerical value, e.g. the activity of a compound in a screen. The user can then select a palette from the ColorBrewer2 color schemes (36). The palette is used to convert the numerical value into a color, which is then used to highlight the compounds in the network with a colored halo (Figure 3). It is also possible to directly specify colors (in standard hexadecimal notation).

USE CASES

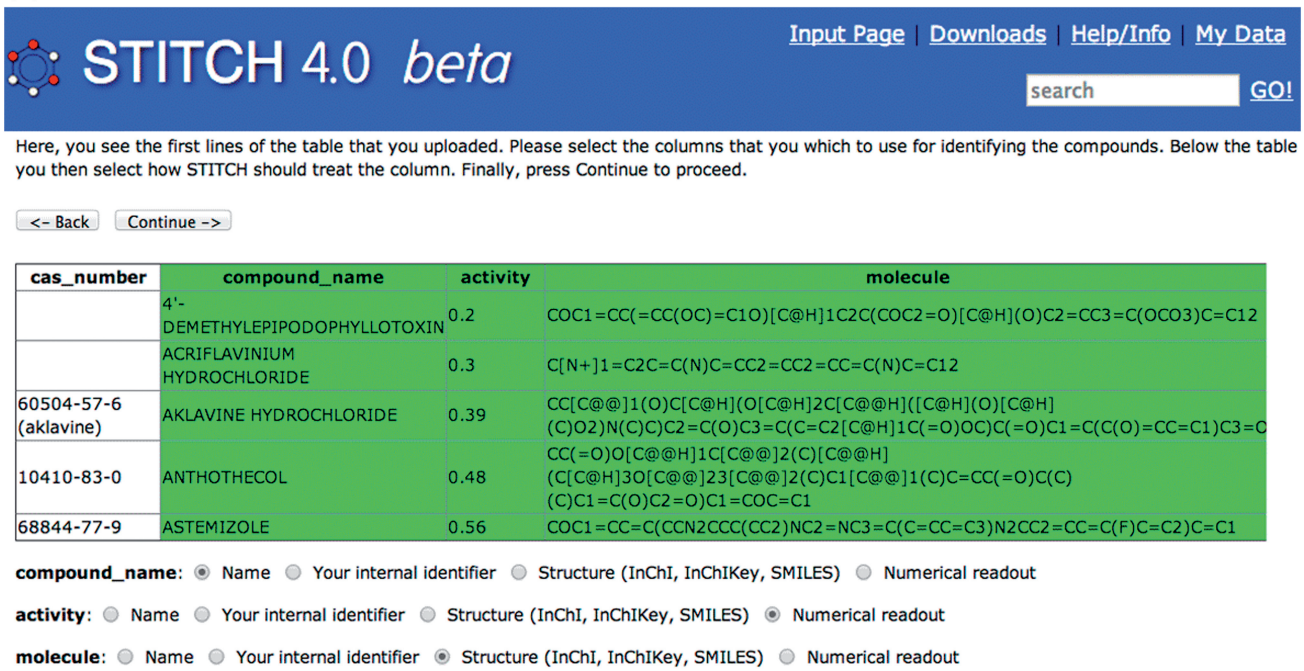
The majority of users access STITCH via the web interface, where networks can be retrieved using single or multiple names of proteins or chemicals. Furthermore, users can query STITCH with protein sequences and chemical structures (in the form of SMILES strings). The networks can then be explored interactively or saved in different formats, including publication-quality images. Proteins and chemicals can be clustered in the

(a)



	A	B	C	D
1	cas_number	compound_name	activity	molecule
2		4'-DEMETHYLEPIPODOPHYLLOTOXIN	0.20	COC1=CC(=CC(OC)=C1O)[C@H]1C2C(COC2=
3		ACRIFLAVINIUM HYDROCHLORIDE	0.30	C[N+]1=C2C=C(N)C=CC2=CC2=CC=C(N)C=C1
4	60504-57-6 (aklavine)	AKLAVINE HYDROCHLORIDE	0.39	CC[C@@]1(O)C[C@H](O[C@H]2C[C@@H]([
5	10410-83-0	ANTHOTHECOL	0.48	CC(=O)O[C@@H]1C[C@@]2(C)[C@@H](C[C
6	68844-77-9	ASTEMIZOLE	0.56	COC1=CC=C(CCN2CCC(CC2)NC2=NC3=C(C=C
7	474-07-7	BRAZILIN	0.64	OC1=CC=C2C3C4=C(CC3(O)COC2=C1)C=C(O
8	56-25-7	CANTHARIDIN	0.72	C[C@]12C3CCC(O3)[C@@]1(C)C(=O)OC2=O
9	1254-85-9	CEDRELONE	0.78	C[C@@]12CCC3[C@@]4(C)C=CC(=O)C(C)(C)
10	34157-83-0	CELASTROL	0.84	[H][C@@]12C[C@@](C)(CC[C@]1(C)CC[C@
11	57-09-0, 6899-10-1 [cet	CETRIMONIUM BROMIDE	0.89	CCCCCCCCCCCCCCCC[N+](C)(C)C
12	6004-24-6, 123-03-5 [an	CETYLPYRIDINIUM CHLORIDE	0.93	CCCCCCCCCCCCCCCC[N+]1=CC=CC=C1
13	3697-42-5, 55-56-1 [chl	CHLORHEXIDINE	0.96	ClC1=CC=C(NC(=N)NC(=N)NCCCCCCNC(=N)N
14	477-27-0	COLCHICEINE	0.99	COC1=CC2=C(C(OC)=C1OC)C1=CC=C(O)C(=O
15	64-86-8	COLCHICINE	1.00	COC1=CC2=C(C(OC)=C1OC)C1=CC=C(OC)C(=
16	28068-69-1	CRASSIN ACETATE	1.00	CC(=O)OC1C\C(C)=C\CC\C(C)=C\CC[C@](C)(
17	50-76-0	DACTINOMYCIN	0.99	CC(C)C1NC(=O)C(NC(=O)C2=C3N=C4C(OC3=
18	71-63-6	DIGITOXIN	0.97	[H][C@]12CCC3C(CC[C@]4(C)[C@H](CC[C@

(b)



cas_number	compound_name	activity	molecule
	4'-DEMETHYLEPIPODOPHYLLOTOXIN	0.2	COC1=CC(=CC(OC)=C1O)[C@H]1C2C(COC2=O)[C@H](O)C2=CC3=C(OCO3)C=C12
	ACRIFLAVINIUM HYDROCHLORIDE	0.3	C[N+]1=C2C=C(N)C=CC2=CC2=CC=C(N)C=C12
60504-57-6 (aklavine)	AKLAVINE HYDROCHLORIDE	0.39	CC[C@@]1(O)C[C@H](O[C@H]2C[C@@H]([C@H](O)[C@H](C)O2)N(C)C)C2=C(O)C3=C(C=C2[C@H]1C(=O)OC)C(=O)C1=C(C(O)=CC=C1)C3=O
10410-83-0	ANTHOTHECOL	0.48	CC(=O)O[C@@H]1C[C@@]2(C)[C@@H](C[C@H]3O[C@@]23[C@@]2(C)C1[C@@]1(C)C=CC(=O)C(C)(C)C1=C(O)C2=O)C1=COC=C1
68844-77-9	ASTEMIZOLE	0.56	COC1=CC=C(CCN2CCC(CC2)NC2=NC3=C(C=CC=C3)N2CC2=CC=C(F)C=C2)C=C1

Figure 2. Data upload. The user can use the batch import form to upload a spreadsheet, e.g. from Microsoft Excel (**a**). STITCH will then show the first five rows of the spreadsheet and ask the user to identify columns that contain the name, chemical structure or a numerical readout (**b**). Selected columns are highlighted in green. STITCH uses heuristics to suggest which kind of information the columns contain, e.g. by identifying SMILES strings as structural descriptors.

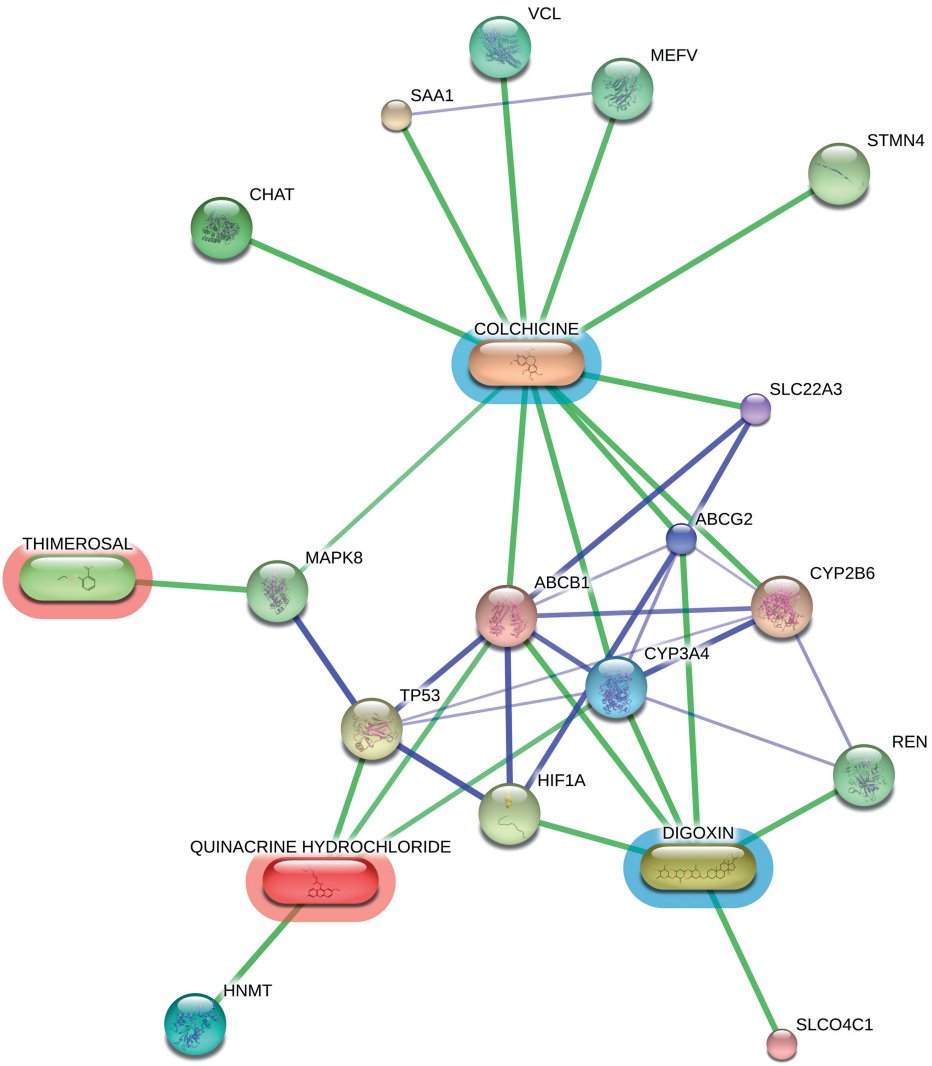


Figure 3. User data and the STITCH network. For four compounds that are part of the example data set from Figure 2, interacting proteins are shown. The numerical readout has been converted to a color on a red–blue gradient. Instead of the normal chemical names used by STITCH, the full names provided in the data set are used, enabling the user to easily recognize the studied chemicals.

interactive network viewer and enriched GO terms among the proteins can be computed (5,37). The set of all interactions is also available for download under Creative Commons licenses (with separate commercial licensing for a subset). In this way, STITCH can be used to drive large-scale studies. Many research groups have already used STITCH 3 in this way; a few examples illustrating different utilities follow: STITCH has been used to determine which proteins cause side effects during drug treatment (38,39) by combining the STITCH network with data from a side effect database (40). The database has also been instrumental for the identification of druggable proteins to predict polypharmacological treatment of diseases on the basis of network topology features (41). For a method that predicts drug targets based on chemogenetic assays in yeast, STITCH has been chosen as a benchmark set (42). Lastly, STITCH has also been integrated into other tools, for example ResponseNet2.0 and QuantMap (43,44).

ACKNOWLEDGEMENTS

The authors wish to thank Yan P. Yuan (EMBL) for his outstanding support with the STITCH servers.

FUNDING

Deutsche Forschungsgemeinschaft [DFG KU 2796/2-1 to M.K.]; Novo Nordisk Foundation Center for Protein Research. Funding for open access charge: European Molecular Biology Laboratory.

Conflict of interest statement. None declared.

REFERENCES

1. Barabási,A.L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–113.

2. Kuhn,M., von Mering,C., Campillos,M., Jensen,L.J. and Bork,P. (2008) STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res.*, **36**, D684–D688.
3. Kuhn,M., Szklarczyk,D., Franceschini,A., Campillos,M., von Mering,C., Jensen,L.J., Beyer,A. and Bork,P. (2010) STITCH 2: an interaction network database for small molecules and proteins. *Nucleic Acids Res.*, **38**, D552–D556.
4. Kuhn,M., Szklarczyk,D., Franceschini,A., von Mering,C., Jensen,L.J. and Bork,P. (2012) STITCH 3: zooming in on protein-chemical interactions. *Nucleic Acids Res.*, **40**, D876–D880.
5. Franceschini,A., Szklarczyk,D., Frankild,S., Kuhn,M., Simonovic,M., Roth,A., Lin,J., Minguez,P., Bork,P., von Mering,C. *et al.* (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.*, **41**, D808–D815.
6. Wang,Y., Xiao,J., Suzek,T.O., Zhang,J., Wang,J. and Bryant,S.H. (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.*, **37**, W623–W33.
7. Gaulton,A., Bellis,L.J., Bento,A.P., Chambers,J., Davies,M., Hersey,A., Light,Y., McGlinchey,S., Michalovich,D., Al-Lazikani,B. *et al.* (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, **40**, D1100–D1107.
8. Roth,B.L., Lopez,E., Patel,S. and Kroeze,W. (2000) The multiplicity of serotonin receptors: uselessly diverse molecules or an embarrassment of riches? *Neuroscientist*, **6**, 252–262.
9. Rose,P.W., Beran,B., Bi,C., Bluhm,W.F., Dimitropoulos,D., Goodsell,D.S., Prlic,A., Quesada,M., Quinn,G.B., Westbrook,J.D. *et al.* (2011) The RCSB protein data bank: redesigned web site and web services. *Nucleic Acids Res.*, **39**, D392–D401.
10. Anastassiadis,T., Deacon,S.W., Devarajan,K., Ma,H. and Peterson,J.R. (2011) Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. *Nat. Biotechnol*, **29**, 1039–1045.
11. Davis,M.I., Hunt,J.P., Herrgard,S., Ciceri,P., Wodicka,L.M., Pallares,G., Hocker,M., Treiber,D.K. and Zarrinkar,P.P. (2011) Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.*, **29**, 1046–1051.
12. Knox,C., Law,V., Jewison,T., Liu,P., Ly,S., Frolkis,A., Pon,A., Banco,K., Mak,C., Neveu,V. *et al.* (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.*, **39**, D1035–D1041.
13. Okuno,Y., Yang,J., Taneishi,K., Yabuuchi,H. and Tsujimoto,G. (2006) GLIDA: GPCR-ligand database for chemical genomic drug discovery. *Nucleic Acids Res.*, **34**, D673–D677.
14. Günther,S., Kuhn,M., Dunkel,M., Campillos,M., Senger,C., Petsalaki,E., Ahmed,J., Urdiales,E.G., Gewiess,A., Jensen,L.J. *et al.* (2008) SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Res.*, **36**, D919–D922.
15. Zhu,F., Shi,Z., Qin,C., Tao,L., Liu,X., Xu,F., Zhang,L., Song,Y., Liu,X., Zhang,J. *et al.* (2012) Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic Acids Res.*, **40**, D1128–D1136.
16. Davis,A.P., Murphy,C.G., Johnson,R., Lay,J.M., Lennon-Hopkins,K., Saraceni-Richards,C., Sciaky,D., King,B.L., Rosenstein,M.C., Wiegers,T.C. *et al.* (2013) The comparative toxicogenomics database: update 2013. *Nucleic Acids Res.*, **41**, D1104–D1114.
17. Kanehisa,M., Goto,S., Sato,Y., Furumichi,M. and Tanabe,M. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.
18. Schaefer,C.F., Anthony,K., Krupa,S., Buchoff,J., Day,M., Hannay,T. and Buetow,K.H. (2009) PID: the pathway interaction database. *Nucleic Acids Res.*, **37**, D674–D679.
19. Croft,D., O'Kelly,G., Wu,G., Haw,R., Gillespie,M., Matthews,L., Caudy,M., Garapati,P., Gopinath,G., Jassal,B. *et al.* (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, **39**, D691–D697.
20. Caspi,R., Altman,T., Dale,J.M., Dreher,K., Fulcher,C.A., Gilham,F., Kaipa,P., Karthikeyan,A.S., Kothari,A., Krummenacker,M. *et al.* (2010) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **38**, D473–D479.
21. Saric,J., Jensen,L.J., Ouzounova,R., Rojas,I. and Bork,P. (2006) Extraction of regulatory gene/protein networks from Medline. *Bioinformatics*, **22**, 645–650.
22. Jensen,L.J., Saric,J. and Bork,P. (2006) Literature mining for the biologist: from information retrieval to biological discovery. *Nat. Rev. Genet.*, **7**, 119–129.
23. Amberger,J., Bocchini,C.A., Scott,A.F. and Hamosh,A. (2009) McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.*, **37**, D793–D796.
24. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.*, **25**, 25–29.
25. Lounkine,E., Keiser,M.J., Whitebread,S., Mikhailov,D., Hamon,J., Jenkins,J.L., Lavan,P., Weber,E., Doak,A.K., Côté,S. *et al.* (2012) Large-scale prediction and testing of drug activity on side-effect targets. *Nature*, **486**, 361–367.
26. Besnard,J., Ruda,G.F., Setola,V., Abecassis,K., Rodriguiz,R.M., Huang,X.P., Norval,S., Sassano,M.F., Shin,A.I., Webster,L.A. *et al.* (2012) Automated design of ligands to polypharmacological profiles. *Nature*, **492**, 215–220.
27. Paolini,G.V., Shapland,R.H., van Hoorn,W.P., Mason,J.S. and Hopkins,A.L. (2006) Global mapping of pharmacological space. *Nat. Biotechnol.*, **24**, 805–815.
28. Rognan,D. (2013) Towards the next generation of computational chemogenomics tools. *Mol. Inf.*, **32**, 1029–1034.
29. Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.
30. Chen,B., Sheridan,R.P., Hornak,V. and Voigt,J.H. (2012) Comparison of random forest and pipeline pilot naïve bayes in prospective QSAR predictions. *J. Chem. Inf. Model*, **52**, 792–803.
31. Hobohm,U., Scharf,M., Schneider,R. and Sander,C. (1992) Selection of representative protein data sets. *Protein Sci.*, **1**, 409–417.
32. Steinbeck,C., Hoppe,C., Kuhn,S., Floris,M., Guha,R. and Willighagen,E.L. (2006) Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics. *Curr. Pharm. Des.*, **12**, 2111–2120.
33. Steinbeck,C., Han,Y., Kuhn,S., Horlacher,O., Luttmann,E. and Willighagen,E. (2003) The Chemistry Development Kit (CDK): an open-source java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.*, **43**, 493–500.
34. Iskar,M., Campillos,M., Kuhn,M., Jensen,L.J., van Noort,V. and Bork,P. (2010) Drug-induced regulation of target expression. *PLoS Comput. Biol.*, **6**, e1000925.
35. Grever,M.R., Schepartz,S.A. and Chabner,B.A. (1992) The National Cancer Institute: cancer drug discovery and development program. *Semin. Oncol.*, **19**, 622–638.
36. Harrower,M. and Brewer,C.A. (2003) ColorBrewer.org: an online tool for selecting colour schemes for maps. *Cartogr. J.*, **40**, 27–37.
37. Szklarczyk,D., Franceschini,A., Kuhn,M., Simonovic,M., Roth,A., Minguez,P., Doerks,T., Stark,M., Muller,J., Bork,P. *et al.* (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.*, **39**, D561–D568.
38. Duran-Frigola,M. and Aloy,P. (2013) Analysis of chemical and biological features yields mechanistic insights into drug side effects. *Chem. Biol.*, **20**, 594–603.
39. Kuhn,M., Al Banchaabouchi,M., Campillos,M., Jensen,L.J., Gross,C., Gavin,A.C. and Bork,P. (2013) Systematic identification of proteins that elicit drug side effects. *Mol. Syst. Biol.*, **9**, 663.
40. Kuhn,M., Campillos,M., Letunic,I., Jensen,L.J. and Bork,P. (2010) A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol.*, **6**, 343.
41. Vitali,F., Mulas,F., Marini,P. and Bellazzi,R. (2013) Network-based target ranking for polypharmacological therapies. *J. Biomed. Inf.*, **46**, 876–881.

42. Heiskanen,M.A. and Aittokallio,T. (2013) Predicting drug-target interactions through integrative analysis of chemogenetic assays in yeast. *Mol. Biosyst.*, **9**, 768–779.
43. Basha,O., Tirman,S., Eluk,A. and Yeger-Lotem,E. (2013) ResponseNet2.0: revealing signaling and regulatory pathways connecting your proteins and genes—now with human data. *Nucleic Acids Res.*, **41**, W198–W203.
44. Schaal,W., Hammerling,U., Gustafsson,M.G. and Spjuth,O. (2013) Automated QuantMap for rapid quantitative molecular network topology analysis. *Bioinformatics*, **29**, 2369–2370.