Rapid #: -10511869

CROSS REF ID: **1319836**

LENDER: **HNK :: Main Library**

BORROWER: **WAU :: Suzzallo Library**

TYPE:	Article CC:CCG
JOURNAL TITLE:	Systems and synthetic biology
USER JOURNAL TITLE:	Systems and synthetic biology,Systems and synthetic biology,
ARTICLE TITLE:	Cheminformatics models based on machine learning approaches for design of USP1/UAF1 abrogators as anticancer agents,,
ARTICLE AUTHOR:	Wahi, Divya
VOLUME:	9
ISSUE:	1-2
MONTH:	
YEAR:	2015
PAGES:	33 -
ISSN:	1872-5325
OCLC #:	

Processed by RapidX: 4/14/2016 8:17:59 PM

ARTICLE SLIP

Borrower: RAPID:WAU

Lending String:

Patron:

Journal Title: Systems and synthetic biology

Volume: 9 **Issue:** 1-2
Month/Year: 2015
Pages: 33 - 43

Article Author: Wahi, Divya

Article Title: Cheminformatics models based on machine learning approaches for design of USP1/UAF1 abrogators as anticancer agents,,

Imprint:

ILL Number: -10511869

Call #: QH313 .S973
Location: ej

ILLiad TN: 216393

Charge
Maxcost:

Shipping Address:
NEW: Suzzallo Library

Fax:
Ariel:

Syst Synth Biol (2015) 9:33–43
DOI 10.1007/s11693-015-9162-1

RESEARCH ARTICLE

Cheminformatics models based on machine learning approaches for design of USP1/UAF1 abrogators as anticancer agents

Divya Wahi · Salma Jamal · Sukriti Goyal · Aditi Singh · Ritu Jain · Preeti Rana · Abhinav Grover

Received: 8 October 2014 / Revised: 14 January 2015 / Accepted: 23 January 2015 / Published online: 30 January 2015

Abstract Cancer cells have upregulated DNA repair mechanisms, enabling them survive DNA damage induced during repeated rapid cell divisions and targeted chemotherapeutic treatments. Cancer cell proliferation and survival targeting via inhibition of DNA repair pathways is currently a very promiscuous anti-tumor approach. The deubiquitinating enzyme, USP1 is known to promote DNA repair via complexing with UAF1. The USP1/UAF1 complex is responsible for regulating DNA break repair pathways such as trans-lesion synthesis pathway, Fanconi anemia pathway and homologous recombination. Thus, USP1/UAF1 inhibition poses as an efficient anti-cancer strategy. The recently made available high throughput screen data for anti USP1/UAF1 activity prompted us to compute bioactivity predictive models that could help in screening for potential USP1/UAF1 inhibitors having anticancer properties. The current study utilizes publicly available high throughput screen data set of chemical compounds evaluated for their potential USP1/UAF1 inhibitory effect. A machine learning approach was devised for generation of computational models that could predict for potential anti USP1/UAF1 biological activity of novel anticancer compounds. Additional efficacy of active compounds was screened by applying SMARTS filter to eliminate molecules with non-drug like features. The structural fragment analysis was further performed to explore structural properties of the molecules. We demonstrated that modern machine learning approaches could be efficiently employed in building predictive computational models and their predictive performance is statistically accurate. The structure fragment analysis revealed the structures that could play an important role in identification of USP1/UAF1 inhibitors.

Keywords USP1 · UAF1 · Inhibitors · Machine learning · Model · Cancer · Anticancer · Cheminformatics · DNA repair

Introduction

Ubiquitin, the regulatory molecule that is central to protein biology, is a small protein, that acts as a controller of cellular life and death by being primarily involved in phenomena such as trafficking, chromatin modelling, DNA repair etc. (Chen and Sun 2009). Ubiquitin modification, namely ubiquitination and deubiquitination, mediated signalling forms a crucial step in a plethora of molecular pathways occurring in living cells and cellular homeostasis (Garcia-Santisteban et al. 2013). While the process of Ubiquitin addition is well characterised and its crucial role in protein biology is established, the essence of Ubiquitin removal still remains unexplored and still evolving (Liang et al. 2014).

Deubiquitinases (DUBs) are a class of enzymes that facilitate the cellular process of ubiquitin removal and

Electronic supplementary material The online version of this article (doi:10.1007/s11693-015-9162-1) contains supplementary material, which is available to authorized users.

D. Wahi · R. Jain · P. Rana · A. Grover (✉)
School of Biotechnology, Jawaharlal Nehru University, New Delhi 110067, India
e-mail: abhinavgr@gmail.com; agrover@jnu.ac.in

S. Jamal · S. Goyal
Department of Bioscience and Biotechnology, Banasthali University, Tonk 304022, Rajasthan, India

A. Singh
Department of Biotechnology, TERI University, Plot No. 10, Institutional Area, Vasant Kunj, New Delhi 110 070, India

serve the main functions of ubiquitin chain editing and ubiquitin rescue from degradation marked proteins (Amerik and Hochstrasser 2004; Nijman et al. 2005b). Apart from the well elucidated degradative roles, emerging roles have highlighted that deubiquitinases are central in function for regulation of other pathways, and is particularly germane to cancer and reversible ubiquitination influences the stability of important oncogenes (Sacco et al. 2010). Mounting evidences in recent, support the view that deubiquitinases be characterised as oncogenes, owing to their involvement in regulating cancer-relevant pathways of tumor development and malignant transformations (Fraile et al. 2012; Hussain et al. 2009). It is increasingly becoming well stated, that DNA damage repair is tightly regulated under ubiquitination and deubiquitination, thus making the concept of DUBs as eminently druggable target more relevant (Hofmann 2009; Ulrich and Walden 2010). There is an interesting subset of DUBs, that have been shown to play a critical role in DNA damage repair responses particularly in cancers (Garcia-Santisteban et al. 2013). Thus, targeting the phenomenon of deubiquitination, as a part of synthetic lethal approach to destroy cancer cells dependent on a compensatory DNA repair pathway for survival due to constant oxidative and replicative stress, has spurred interests to use it as an anti-cancer strategy (Curtin 2012; Hoeller and Dikic 2009).

DNA damage can occur via single-stranded breaks (SSBs), double-stranded breaks (DSBs) or formation of DNA adducts, due to covalent modification or base cross-linking (Huang and D'Andrea 2006). Thus, making DNA repair a critical phenomenon responsible to maintain cellular genome integrity and any disruption in these repair mechanisms could render cells prone to damage, death and mutagenicity (Murai et al. 2011). Since cancer cells are rapidly proliferating, spontaneous DNA breaks or increased reactive oxygen species (ROS) induce base oxidation or exposure to Chemotherapeutics, as in case of Cancer treatment; causes their DNA to commonly exhibit increased DNA mutation rates and greater genomic instability (Branzei and Foiani 2008; Kennedy and D'Andrea 2006). This renders cancer cells, victim to frequent DNA damage and reliant on an activated DNA damage repair mechanism for survival (Helleday et al. 2008).

Cancer cells frequently develop proliferative defects in DNA and upregulate one of the following major DNA repair pathways, namely: homologous recombination (HR) repair, mismatch repair (MMR), base excision repair (BER), nucleotide excision repair (NER), nonhomologous end-joining (NHEJ) or translesion DNA synthesis (TLS)(Kennedy and D'Andrea 2006; Murai et al. 2011). Although the exact mechanism used in repair pathway selection repair activity coordination is not known, but recent studies highlight the importance of ubiquitin in DNA repair regulation (Murai et al. 2011). This critical need of highly active repair in constantly dividing carcinoma cells can be exploited to target cancer cells by using inhibitors against proteins, such as ubiquitinases and deubiquitinases, essential for the repair pathway functioning.

Deubiquitinating enzymes (DUBs) cleave the isopeptide bond between the ubiquitin C-terminal carboxylate and lysine side-chain amino group on target proteins. As more than 100 different types of DUBs are known, they are classified into five major families and the Ubiquitin specific proteases (USPs) among them, constitute as the largest family (Nijman et al. 2005b). The occurrence of Deregulated USP1 in various types of carcinoma makes it the most promiscuous pharmacological intervention target of the USPs family, and thus USP1 holds a special attraction. USP1 is amongst the most well characterised deubiquitinases which have been found to play critical role in DNA damage repair mechanisms and more. Previous studies have validated that USP1 disruption results in heightened sensitivity to DNA crosslinking agents in chicken (Oestergaard et al. 2007) and USP1 knockout results in mitomycin C hypersensitivity in murine model (Kim et al. 2009). Also, USP1 has been associated with the function of deubiquitinating PCNA in Trans-lesion synthesis pathway (Huang et al. 2006) and FANCD2/FANCI in Fanconi anemia pathway (Nijman et al. 2005a), hampering the respective DNA damage avoidance mechanisms. PCNA (Proliferating Cell Nuclear Antigen) monoubiquitination is critical to the Translesion DNA damage repair response (TLS), that promotes lesion bypass at the site of stalled replication fork due to a damaged base(Kannouche et al. 2004; Zhuang et al. 2008). USP1 mediated reversion of PCNA monoubiquitination prevents untimely TLS polymerase recruitment, contributing to the maintenance of genome stability (Jones et al. 2012). FANCD2 and FANCI monoubiquitination regulates in the DNA Interstrand crosslink repair that occurs during genomic damage in the rare genetic anomaly Fanconi anemia (Kim and D'Andrea 2012; Sims et al. 2007). USP1 is known to deubiquitinate FANCD2 and FANCI for crucially maintaining the FA pathway functioning, and USP1 knockouts recapitulate FA phenotype (Kim et al. 2009; Oestergaard et al. 2007).

USP1 participates in the homologous recombination during double stranded break via suppression of non-homologous end joining repair pathway (Murai et al. 2011). USP1 inhibition has further been shown to target the degradation of ID (inhibitor of DNA binding 1) proteins (Mistry et al. 2013), which are known to be overexpressed in various cancer types (Fong et al. 2004). In addition it has been shown, that USP1 inhibition is a potential differentiation therapy target, since USP1 mediated deubiquitination of ID proteins preserves tumorigenic capacity and "stemness" in osteosarcoma (Williams et al. 2011). Thus,

making USP1 inhibition a potent anti-cancer strategy. It has been uncovered, that all these ubiquitin addition and removal regulation is carried out by USP1 in a heterodimeric complex form, associated with its cofactor UAF1 (USP1 associated factor 1), which modulates the enzyme activity (Cohn et al. 2007, 2009). UAF1 binding induces conformational changes in USP1 active site increasing the enzyme activity dramatically by stabilizing it (Cohn et al. 2007; Villamil et al. 2012) and plays critical role as a cofactor. It is to be noted that since DNA damaging chemotherapeutics make cancer cells reliant on DNA repair pathways for survival, USP1 inhibitors could act as efficacious anti-cancer molecules, when used in combination them. Also, since therapeutic intervention with an USP1 inhibitor would be a single agent therapy, it would be leading to better efficacy, likely fewer side effects and could be more tumor selective. Thus, this vouches USP1/UAF1 complex an interesting target with a lot of scope to be explored as an anti-cancer target.

The present study aims to employ in silico methods for drug development (Dhanjal et al. 2014; Goyal et al. 2014). In this study we explore the predictive capacities of machine learning for creating classification models from high throughput screens which could be explored to identify compounds having therapeutic anti tumor effects. We started with the Cheminformatics study by building models for predictive classification from high throughput assays targeted against USP1/UAF1. Additional analysis implies the application of Substructure Fragments, to study enriched substructure fragments which could further be useful in screening more inhibitors to USP1/UAF1. Our study reveals accurate and efficient predictive computational models that could screen for molecules with high potential to being modelled as anti-tumor compounds. Thus, we demonstrate that such machine learning based predictive modelling, data mining and screening has a lot of potential and could be effectively used to identify molecules for the drug discovery process.

Materials and methods

Bioassay data set

In the current study, the assay corresponding to assay identifier AID 743255 was used that aimed at screening NIH Molecular Libraries Small Molecule Repository (MLSMR) to identify small molecules that can act as inhibitors against USP1/UAF1 complex (Liang et al. 2014). The assay was obtained from PubChem (Wang et al. 2012), a database primarily providing information about the biological activities of small molecules, maintained by National Center for Biotechnology Information (NCBI).

The assay is a Public domain Molecular Data that comprises of a miniaturized quantitative high-throughput screen (qHTS) that monitors activity of compounds against USP1/UAF1 using an ubiquitin–rhodamine110 substrate (Liang et al. 2014). The dataset AID 743255 consisted of a total of 389,560 compounds tested for inhibitory effect against USP1. Compounds were characterised on the basis of their PubChem Activity score, where a score between 40 and 100 accounted for active compounds (n = 904) and a score of 0 accounted for inactive compounds (n = 369,838). Also, compounds with an activity score of 1–39 were marked as Inconclusive and others marked as Unspecified, were not considered further in our analysis due to the probable uncertainty in their predictive ability of bioactivity for our models.

Dataset pre-processing and generation of molecular descriptors

The chemical structures of the active and inactive molecules of the AID 743255 were downloaded from PubChem in the structural data format (SDF). Owing to the large size of the dataset containing all the compounds, it was not possible to process it further as a single file. This prompted us to split the SDF files obtained above, into smaller files using the SplitSDFiles Perl Script available at Mayachem tools (Sud 2010). These structures were further imported into the freely available software, PowerMV, which is primarily used for statistical analysis, descriptor generation and molecular viewing (Liu et al. 2005). PowerMV transforms the information encoded in chemical structures into numbers which are termed as descriptors. Using PowerMV, 2D Molecular descriptors were generated for the actives and inactives. PowerMV generates in total 179 descriptors for all the input compounds, of which 147 descriptors account for pharmacophore Fingerprints, 24 descriptors account for Weighted Burden numbers and 8 for property descriptors. Pharmacophore fingerprints are binary descriptors based on bioisosteric principles according to which two atoms or groups having same biological activity are assigned the same class. Pharmacophore based descriptors are divided into six categories which include negatively and positively charged atoms or groups, hydrogen bond donors and acceptors and ring systems containing aromatic and hydrophobic centers. Another class of descriptors generated by PowerMV includes weighted burden numbers which are continuous descriptors based on a connectivity matrix, the Burden connectivity matrix takes into account three properties, electronegativity, partial charge and atomic lipophilicity. The property set of descriptors generated by PowerMV include eight properties, partition coefficient, polar surface area, number of rotatable bonds, H-bond donors and acceptors, molecular

weight, blood–brain indicator and bad group indicator (Liu et al. 2005). The dimensionality of the data was reduced by removing redundant descriptors across the entire dataset. Further using Perl script, the data set was split into 20 % independent Test set and 80 % Training cum validation set across five folds, employed during the entire study.

Best first descriptors

Feature selection techniques have been widely used to reduce the dimensionality of the dataset, computation time involved as well as to reduce the noise from the data. Most traditional models have been constructed using all the available descriptors or attributes of the data. But as recognized earlier (Dudek et al. 2006), we sought to explore if amongst the large number of descriptors, only few were actually of relevance in contributing towards an efficient model building. And identification of such a set of few descriptors, which were enough to characterize a model as the best first descriptors. The key rationale for such search for descriptors was to make the model computation time efficient and reduce the subsequent noise that eventually affects the end model quality. The feature selection algorithms search throughout the dataset for all possible combinations of features and put forward a subset of features contributing most towards the classification. Feature selection techniques use an attribute evaluator method in combination with a search method. In the present study, we have used CfsSubsetEval module in combination with the BestFirst search method implemented in Weka. CfsSubsetEval evaluator method considers the predictive ability of every single feature and looks for subset of features having high correlation with the predicting class.

Machine learning implementation: classification algorithms

The science of machine learning revolves around construction of computational models and algorithms for predictive learning based on the classifiers studied across the training datasets (Melville et al. 2009). Cheminformatics employs the machine learning techniques to efficiently build predictive models from sets of known compounds and compute molecular properties and biological activities for unknown compounds. In most popular approaches, molecular dataset is classified on the basis of activity (as actives and inactives) and a Binary classification based on the molecular descriptors is performed. Earlier studies have shown to accurately predict various other datasets having active molecules for malaria (Jamal et al. 2013), tuberculosis (Periwal et al. 2011, 2012), leishmania (Jamal and Scaria 2013) etc. Also, multiple class based classification has been extensively attempted and reviewed (Jensen et al. 2006).

Thus, machine learning algorithms and their implementation techniques are numerous and widely varied. In this study, we have employed the most popular four algorithms, which have been labelled as both computation time-efficient and classification-accurate, viz. Naïve Bayes, Random forest, J48 and SMO. The Bayesian theorem based classifier Naïve Bayes, is the simplest of all probabilistic classifiers. It classifies based on the assumption that, each prediction based on a descriptor is independent of the other descriptors, where the final prediction is a product of all the descriptor based probabilities. The Random forest algorithm on the other hand is a decision tree based classifier, where each independently constructed tree node is split based on the basis of the best predictor subset randomly chosen at it. It is noted to be the best available and time effective classifier with the most accurate classification results. The J48 classifier builds binary decision trees based on the information gain obtained by splitting the data into smaller subsets, each time giving idea of importance of an attribute in the dataset that can be used to make a decision. The SMO algorithm is primarily employed for Support vector machine training where it breaks the arising quadratic programming problems into a series of smaller problem subsets, further solved analytically. These algorithms have already been discussed extensively in (Breiman 2001; Cortes 1995; Friedman et al. 1997; Quinlan 1993) respectively.

Stringent algorithms: cost sensitive classification

The machine learning technique of the current study uses classification models to create binary bioactivity classification model of actives and inactives. This is done keeping in mind a key feature of unbiased virtual high-throughput bioassay data screening, i.e., high imbalance in the number of actives and inactives (class imbalance) (Blagus and Lusa 2010). There are two ways of generating classification models, using base classifiers as well as cost sensitive classifiers. Base classifiers use equal costs for all the misclassification errors and thus are not prefered for generating models when there is a high disparity in the dataset. Since the standard classifiers are based on presumption of equal class data (Japkowicz 2000), a cost sensitive classification is employed. Just as misdiagnosing a diseased person is far more serious than stating a healthy person as diseased, loosing active compounds in a dataset can be more costly than including a few inactives. Cost sensitive classification takes misclassification costs into account and the instances are predicted to have class with lowest misclassification cost. A cost sensitive learning involves assigning of a high cost of misclassification on the marginal class, which is then attempted to be reduced by the algorithm. Such a cost sensitive classification system abrogates the

misclassification errors and has been effectively employed previously (Elkan 2001). For any instance x to be classified as belonging to class i or j, the classifier calculates the expected cost of classifying x for all the classes taking into consideration all the possibilities and assigns an instance x the classes which has minimum expected cost.

In the current study, we employed WEKA (Waikato Environment for Knowledge Analysis), a popular collection of machine learning software algorithms, for performing the machine learning, modelling and data mining tasks (Bouckaert et al. 2010). Weka uses algorithms to perform functions such as data pre-processing, classification, clustering, feature selection, visualization and analysis. It can be used to introduce a binary classification based cost sensitivity in the base classifiers, via a 2 × 2 confusion matrix, consisting of the following four sections: true positives (TP) for the active compounds correctly classified as actives; false positives (FP) for the inactive compounds incorrectly classified as actives; true negatives (TN) for the inactive compounds correctly classified as inactives and false negatives (FN) for active compounds incorrectly classified as inactives. Keeping in consideration the criticality of false negative predictions over false positives in the development of classifiers for compound selection experiments, a misclassification cost was set on false negatives. The false negatives were minimized via a serial arbitrary cost value increment to optimize the predictions at the expense of increasing the false positives. Additionally, to constrain the increase in the rate of false positives, we set an empirical upper limit of 20 % on the false positives. Weka does not employ any rules for setting any misclassification cost and the cost is exclusively dependent on the base classifier used (Schierz 2009). The machine learning based computational models were generated using training data and the performance of the models was assessed using the test set. Five-fold cross validation was used during which the training set was randomly divided into five subsets, each time four subsets were used as train set and the remaining set was used as test set. This process was repeated until each subset had been used as test set at least once. Further using the 'supplied test set' option in Weka, the 20 % test cum validation set was supplied and the performance of the generated model was evaluated using various statistical measures.

Statistical chemi-informatic model assessment

Our performance assessment for the classification models was based on the standard machine learning statistical measures such as Sensitivity, Specificity, and Accuracy, Balanced classification Rate (BCR), Receiver operating characteristic curve (ROC) and Matthews Correlation Coefficient (MCC). Sensitivity, Specificity and Accuracy are computed from the True Positive Rate (TPR), False Negative Rate (FNR), True Negative Rate (TNR) and False Positive Rate (FPR). *Sensitivity* or True Positive Rate (TPR) is defined as proportion of actual actives, correctly predicted as Active [TP/(TP + FN)]. Specificity or True Negative Rate (TNR) is defined as proportion of actual inactives, correctly predicted as Inactive [TN/(TN + FP)]. The overall effectiveness of a Binary classifier is assessed by the Accuracy [(TP + TN)/(TP + TN + FP + FN) × 100] and is defined as proportion of true results (both actual actives and actual inactives). G-mean (Geometric Mean) is defined as the measure of central tendency that computes the average of specificity and sensitivity and is denoted by sqrt(sensitivity × specificity). The Receiver Operating Characteristic (ROC) curve is the graphical representation of true positive rate versus false positive rate and the plot illustrates the performance of the binary classifier as the Area under the Curve (AUC). The Matthews Correlation Coefficient (MCC) is defined as the measure that computes the quality of binary classification [(TP × TN)−(FP × FN)]/sqrt [(TP + FP) (TP + FN) (TN + FP) (TN + FN)].

SMARTS filtering

We employed the Smiles Arbitrary Target Specification (SMARTS) filters to our dataset to eliminate all the molecules with typical fragments that render toxicity or reactivity to them for being potential drug compounds, via the online server of SMARTS filter available at http://pasilla.health.unm.edu/tomcat/biocomp/smartsfilter. The web application applies substructure screens from five screening filters namely PAINS, ALARM NMR, Oprea, Blake and Glaxo to prioritize drug likely molecules and removing false positives. The Pan-Assay Interference Compounds (PAINS) screens on the basis of sub-structural features used for promiscuous compound identification from multiple high throughput screens, which were reported active although they may not have the predicted activity (Baell and Holloway 2010). The ALARM NMR is a critical filter in identifying reactive compounds by nuclear magnetic resonance which are likely the reactive false positives and could oxidize or alkylate target proteins (Huth et al. 2005). The Oprea, Blake and Glaxo filters remove compounds with non-drug like properties and comprise of substructure search based filters that perform the filtering on presence of unsuitable leads, reactive functional group or unsuitable natural product (Hann et al. 1999).

Sub-structural fragments analysis

The inhibitors and the non-inhibitors downloaded from PubChem were further mined for the presence of sub-structural motifs that could be critical to its inhibitory

function. The substructure fragment analysis was performed following the methodology of substructure pattern recognition explained by Shen et al. (Cheng et al. 2011; Shen et al. 2010). A dictionary comprising of 307 substructure (SMARTS) patterns called as the substructure fingerprints (SubFP) was used from the free online software resource PaDEL. The patterns obtained were subsequently analysed via the substructure fragment analysis (Jensen et al. 2007). The Fragment Frequency, i.e., the frequency of the presence of a particular fragment in Actives and Inactives was computed with the formula:

$$Fragment\ Frequency = \frac{N_{Fragment\ class} \times N_{total}}{N_{Fragment\ total} \times N_{class}}$$

where $N_{Fragment\ class}$ is denoted by the number of actives containing the SMARTS fragment, N_{total} is denoted by the total number of compounds in the data set, i.e. sum of actives and inactives, $N_{fragment\ total}$ is denoted by the total number of active and inactive compounds containing the SMARTS fragments and N_{class} is denoted by the total number of compounds in a class in the dataset.

Results and discussion

The current study design aims to identify small molecule inhibitors against USP1/UAF1 complex using the freely available dataset from PubChem corresponding to AID 743255. A total of 389,560 compounds were used comprising of 904 actives and 369,838 inactive compounds.

Data sets, molecules and model construction

The active (n = 904) and inactive (n = 369,838) molecules of the data set were downloaded from PubChem and a total of 179 2D molecular descriptors were generated for them using PowerMV. Post data pre-processing (as detailed in "Materials and methods") the number of contributing descriptors came down to 154, indicating around 15 % decrease in descriptor numbers (Supplementary file 1). The experiments were divided into two sets: one set comprised of the standard base classifications and the other consisted of cost-sensitive classification models for the same base classifiers. Since there was an imposition of misclassification cost on false negatives, an upper threshold of below 20 % was set on the false positives. As hypothesised, the misclassification cost introduction caused a significant decrease in the false negative counts and increased the count of true positives indicating towards the increased robustness of the cost sensitive model. A series of models were generated and the results of statistical performance parameters of best models for each classifier in both categories of experiments have been tabulated in Tables 1 and 2. All these statistics were based on the independent test set.

Best first descriptors

The 154 descriptors obtained after applying RemoveUseless descriptors were further filtered using BestFirst attribute selection method in combination with CfsSubsetEval module of Weka. A total of 45 highly relevant descriptors were selected using Weka to reduce the nonspecific features and build a more robust classification model (Supplementary file 1).

Model evaluation

Multiple models were generated using a fivefold cross validation on the training data sets as described in the "Materials and methods". There was a comparison between the Base classifier models (NB, RF, J48, and SMO) and cost sensitive models (cs-NB, cs-RF, cs-J48, cs-SMO). The base models were implemented with different misclassification costs until the threshold of false positives

Table 1 Indicating the classification result for the classification with full attributes

Classifier	TPR	FPR	Accuracy (%)	G mean	ROC area	BCR (%)	MCC	Cost
Naïve Bayes	54.4	19.9	80.06	66.05	75.3	67.29	0.04	28
Random forest	80	19.6	80.35	80.18	87	80.18	0.07	150,000
J48	67.2	19.3	80.65	73.65	77.3	73.96	0.06	23,000
SMO	78.3	20.8	79.19	78.76	84.5	78.76	0.07	450

Table 2 Indicating the classification result for the best first classification

Classifier	TPR	FPR	Accuracy (%)	G mean	ROC area	BCR (%)	MCC	Cost
Naïve Bayes	51.1	19.9	80.01	63.98	72.8	65.60	0.04	102
Random forest	79.4	18.6	81.35	80.40	87.2	80.40	0.08	140,000
J48	70.6	19.9	80.1	75.19	78.3	75.34	0.06	22,500
SMO	61.7	19.7	80.21	70.35	78.7	70.96	0.05	300

was reached and thus, the best cost optimized models were obtained for each classifier (cs-models). Since the computational models were generated using two sets of descriptors, one using 154 descriptors and the other using 45 BestFirst descriptors, the primary evaluation of both sets of models was based on the Sensitivity and Specificity curves (Figs. 1, 2) and the overall classifier efficiency for model generation is based on the accuracy. All the models yielded an accuracy of about 80 %. Other parameters such as AUC and G-mean were also computed to measure model robustness, since Accuracy alone was not enough to assess model performance. Furthermore, we computed the AUC from the ROC plot to further introduce a comprehensive test of model efficiency and robustness using both sets of descriptors and all models had significant AUC on the ROC plots (Figs. 3, 4).

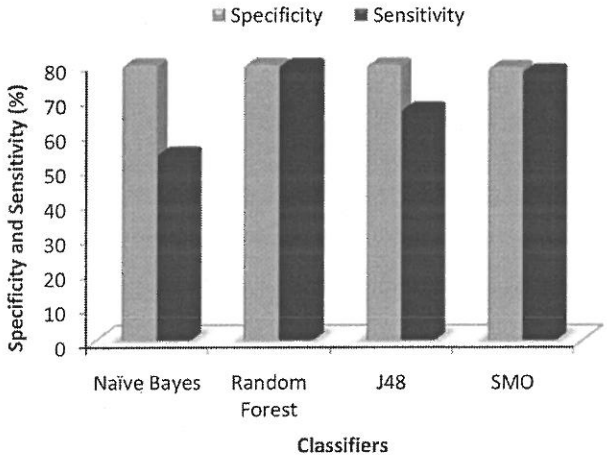

Fig. 1 Plot of comparison between specificity and sensitivity using RemoveUseless (154) attributes

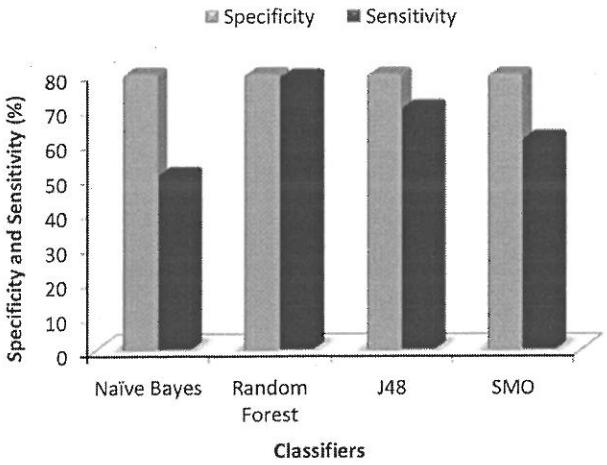

Fig. 2 Plot of comparison between specificity and sensitivity using BestFirst (45) attributes

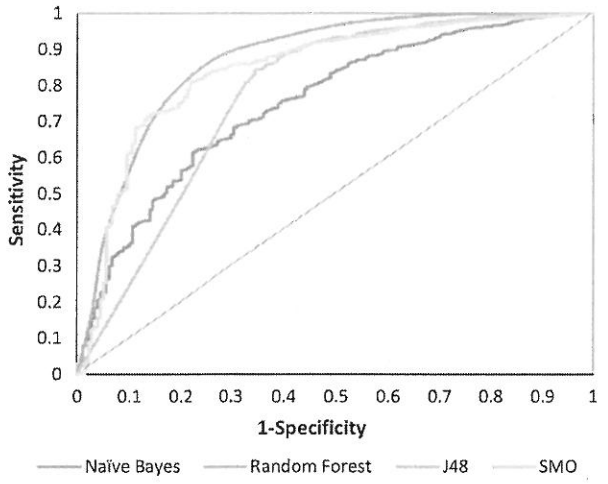

Fig. 3 AUC depicted by the ROC plot for RemoveUseless (154) attributes

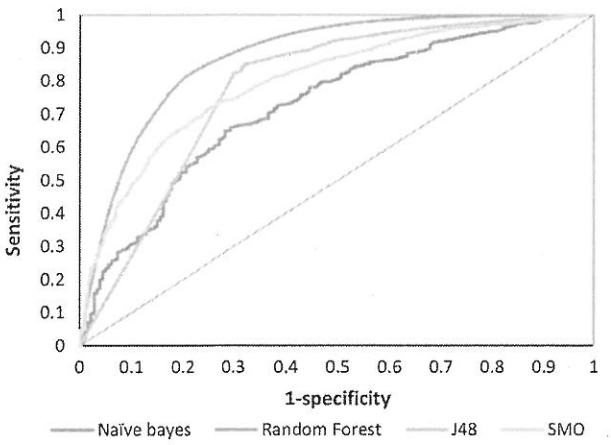

Fig. 4 AUC depicted by the ROC plot for BestFirst attributes

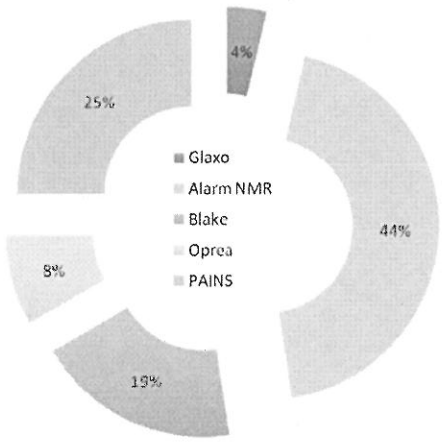

Fig. 5 Pie chart depicting the percentage of molecules that failed each of the SMARTS filter

Table 3 indicating the frequencies of substructure fragments present in the actives and inactives. Substructures in italics are indicative of substructures with p value of less than 0.01, that are characteristic to actives

Substructure Fragment Number	Structure Name	Frequency in actives	Frequency in inactives
SubFP302	Rotatable bond	0.95	1.00
SubFP287	Conjugated double bond	1.27	1.00
SubFP279	Annelated rings	1.77	1.00
SubFP137	Vinylogous ester	1.50	1.00
SubFP300	1,3-Tautomerizable	0.84	1.00
SubFP88	Carboxylic acid derivative	0.79	1.00
SubFP1	Primary carbon	0.91	1.00
SubFP301	1,5-Tautomerizable	1.20	1.00
SubFP135	Vinylogous carbonyl or carboxyl derivative	1.64	1.00
SubFP303	*Michael acceptor*	*4.85*	*0.99*
SubFP23	*Amine*	*2.43*	*1.00*
SubFP49	*Ketone*	*4.06*	*0.99*
SubFP181	Hetero N nonbasic	0.76	1.00
SubFP2	Secondary carbon	0.62	1.00
SubFP138	Vinylogous amide	1.78	1.00
SubFP5	*Alkene*	*3.30*	*0.99*
SubFP18	Alkylarylether	0.66	1.00
SubFP182	Hetero O	1.23	1.00
SubFP177	Oxoarene	1.25	1.00
SubFP98	Amide	0.34	1.00
SubFP76	*Enamine*	*9.43*	*0.98*
SubFP100	Secondary amide	0.32	1.00
SubFP3	Tertiary carbon	0.77	1.00
SubFP169	*Phenol*	*4.54*	*0.99*
SubFP32	*Secondary mixed amine*	*7.55*	*0.98*
SubFP33	*Tertiary mixed amine*	*2.06*	*1.00*
SubFP305	*CH-acidic*	*410.11*	*0.00*
SubFP85	Carboxylic ester	0.79	1.00
SubFP214	Sulfonic derivative	0.65	1.00
SubFP84	Carboxylic acid	2.00	1.00
SubFP171	Arylchloride	0.67	1.00

Among the base classifier models (NB, RF, J48, SMO) and cost sensitive models (cs-NB, cs-RF, cs-J48, cs-SMO), the cost sensitive models performed better. Nevertheless, among the cost sensitive models, the Random Forest model was the best classifier and Naïve Bayes was the least on performance statistics. The random forest model had a BCR of 80.40 %, an accuracy of 81.35 %, sensitivity of 79.44 %, and specificity of 81.36 % and thus it performed better than others giving the overall best classification.

Filtering undesirable structures: SMARTS

An input of 13931 molecules through the SMARTS yielded 9.8 % molecules which passed all the filters, which was 1365 of the total and 12,566 (90.2 %) compounds failed to pass all the components of SMARTS filter. 45 % of the total molecules failed the PAINS screen (6,275 molecules) and 79.1 % of the total failed the ALARM NMR filter (11,018 molecules). But only 15 % of the total molecules failed the Opera filter (2,091 molecules). Similarly 34.2 % (4,769) molecules failed the Blake filter and 6.5 % (911) failed the Glaxo filter. Figure 5 depicts the detailed description of the percentage of molecules that failed each filter. We observed that most of the molecules did not pass through ALARM NMR (79.1 %) filter followed by PAINS filter (45.0 %) and blake filter (34.2 %). The list of 1365 molecules that passed the SMARTS filter is provided as Supplementary file 2.

Analysis of the sub-structure fragments

To initiate a better understanding of the sub-structural details that could be responsible for potential inhibitory effect against USP1/UAF1, we studied active and inactive datasets, searching for occurrence of substructure

fragments using the Substructure fingerprints (SubFP). Substructures with a statistically significant p value of less than 0.01 were considered as contributing and included for analysis. Table 3 indicates the details of the analysed substructure fragments along with a mention of their frequencies in the actives and inactives. CH-acidic, enamine, secondary mixed amine, michael acceptor, phenol, ketone, alkene, amine, tertiary mixed amine turn out to be highly significant and indicative of substructures that are characteristic to actives. Certain other substructure patterns that were characteristic of actives than inactives include carboxylic acid, vinylogous amide, annelated rings, vinylogous carbonyl, vinylogous ester, conjugated double bond, oxoarene, hetero O and 1,5-tautomerizable.

Evaluation of performance of models on external test set

The machine learning models generated in the present study were evaluated for their performance using external dataset of 23165 compounds available from Sigma-Aldrich. The external test set compounds were tested for their activity using four models generated using the Bestfit attributes. The NB, RF, J48 and SMO models predicted 3,913, 5,089, 5,081 and 5,258 compounds as active, respectively. A consensus from all the four models resulted in 981 compounds which could be potential USP1/UAF1 inhibitors.

Conclusion

Targeting cancer via inhibition of USP1/UAF1 is still an emerging and promiscuous domain. Developing a drug that would target DNA damage repair regulated by USP1/UAF1 will be accounted as a single agent therapy. It would aim to have better efficacy, lesser side effects, non-selectivity and more tumor selectivity, than the current conventional anti-cancer therapies. The rational drug discovery process includes experimental screening of large chemical libraries for their bioactivity against target proteins which is very costly as well as time consuming. Computational pre-screening of compounds using cheminformatic approaches has been very effectively used in prioritising compounds in silico and thus accelerating drug discovery process. The open access and free availability to public domain databases containing high throughput chemical screens made it opportune and prompted us with possibilities to do this study. These enormous experimental data from the high throughput screenings (HTS) have been effectively analysed using the data mining and machine learning methods, that build an in silico extension to HTS. They employ construction of predictive models based on the chemical structures and properties of the known compounds. We generated in silico predictive computational models via machine learning approaches to perform virtual screening of the bioassay dataset that was generated for finding out potent USP1/UAF1 inhibitors. We aimed to generate a reliable model that could rapidly and accurately predict the anti USP1/UAF1 activity of compounds, so that genuine hits are optimised with reduced requirements of carrying out cost intensive biological screenings. The current study fairly demonstrates that chemical descriptors based designing and implementation of predictive models from comprehensive machine learning techniques could potentiate low cost large scale virtual screening to up thrust the drug discovery process.

Acknowledgments AG is thankful to Jawaharlal Nehru University for usage of all computational facilities. AG is grateful to University Grants Commission, India for the Faculty Recharge position.

Conflict of interest The authors declare that they have no conflict of interest.

References

Amerik AY, Hochstrasser M (2004) Mechanism and function of deubiquitinating enzymes. Biochim Biophys Acta 1695:189–207

Baell JB, Holloway GA (2010) New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. J Med Chem 53:2719–2740. doi:10.1021/jm901137j

Blagus R, Lusa L (2010) Class prediction for high-dimensional class-imbalanced data. BMC Bioinformatics 11:523. doi:10.1186/1471-2105-11-523

Bouckaert RR, Frank E, Hall MA, Holmes G, Pfahringer B, Reutemann P et al. (2010) Weka—Experiences with a Java Open-Source Project. J Mach Learn Res 10:2533–2541

Branzei D, Foiani M (2008) Regulation of DNA repair throughout the cell cycle. Nat Rev Mol Cell Biol 9:297–308. doi:10.1038/nrm2351

Breiman L (2001) Random forests. Mach Learn 45:5–32. doi:10.1023/a:1010933404324

Chen ZJ, Sun LJ (2009) Nonproteolytic functions of ubiquitin in cell signaling. Mol Cell 33:275–286. doi:10.1016/j.molcel.2009.01.014

Cheng F et al (2011) Classification of cytochrome P450 inhibitors and noninhibitors using combined classifiers. J Chem Inf Model 51:996–1011. doi:10.1021/ci200028n

Cohn MA, Kowal P, Yang K, Haas W, Huang TT, Gygi SP, D'Andrea AD (2007) A UAF1-containing multisubunit protein complex regulates the Fanconi anemia pathway. Mol Cell 28:786–797

Cohn MA, Kee Y, Haas W, Gygi SP, D'Andrea AD (2009) UAF1 is a subunit of multiple deubiquitinating enzyme complexes. J Biol Chem 284:5343–5351. doi:10.1074/jbc.M808430200

Cortes CVV (1995) Support vector networks. Mach Learn 20:273–297

Curtin NJ (2012) DNA repair dysregulation from cancer driver to therapeutic target. Nat Rev Cancer 12:801–817. doi:10.1038/nrc3399

Dhanjal JK, Goyal S, Sharma S, Hamid R, Grover A (2014) Mechanistic insights into mode of action of potent natural

antagonists of BACE-1 for checking Alzheimer's plaque pathology. Biochem Biophys Res Commun 443:1054–1059
Dudek AZ, Arodz T, Galvez J (2006) Computational methods in developing quantitative structure-activity relationships (QSAR): a review. Comb Chem High Throughput Screen 9:213–228
Elkan C (2001) The foundations of cost-sensitive learning. In: Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, vol 2. pp 973–978
Fong S, Debs RJ, Desprez PY (2004) Id genes and proteins as promising targets in cancer therapy. Trends Mol Med 10:387–392. doi:10.1016/j.molmed.2004.06.008
Fraile JM, Quesada V, Rodriguez D, Freije JM, Lopez-Otin C (2012) Deubiquitinases in cancer: new functions and therapeutic options. Oncogene 31:2373–2388. doi:10.1038/onc.2011.443
Friedman N, Geiger D, Goldszmidt M (1997) Bayesian network classifiers. Mach Learn 29:131–163. doi:10.1023/a:1007465528199
Garcia-Santisteban I, Peters GJ, Giovannetti E, Rodriguez JA (2013) USP1 deubiquitinase: cellular functions, regulatory mechanisms and emerging potential as target in cancer therapy. Mol Cancer 12:91. doi:10.1186/1476-4598-12-91
Goyal M, Dhanjal JK, Goyal S, Tyagi C, Hamid R, Grover A (2014) Development of dual inhibitors against Alzheimer's disease using fragment-based QSAR and molecular docking. BioMed Res Int 2014:979606. doi:10.1155/2014/979606
Hann M, Hudson B, Lewell X, Lifely R, Miller L, Ramsden N (1999) Strategic pooling of compounds for high-throughput screening. J Chem Inf Comput Sci 39:897–902
Helleday T, Petermann E, Lundin C, Hodgson B, Sharma RA (2008) DNA repair pathways as targets for cancer therapy. Nat Rev Cancer 8:193–204. doi:10.1038/nrc2342
Hoeller D, Dikic I (2009) Targeting the ubiquitin system in cancer therapy. Nature 458:438–444. doi:10.1038/nature07960
Hofmann K (2009) Ubiquitin-binding domains and their role in the DNA damage response. DNA Repair (Amst) 8:544–556. doi:10.1016/j.dnarep.2009.01.003
Huang TT, D'Andrea AD (2006) Regulation of DNA repair by ubiquitylation. Nat Rev Mol Cell Biol 7:323–334. doi:10.1038/nrm1908
Huang TT et al (2006) Regulation of monoubiquitinated PCNA by DUB autocleavage. Nat Cell Biol 8:339–347. doi:10.1038/ncb1378
Hussain S, Zhang Y, Galardy PJ (2009) DUBs and cancer: the role of deubiquitinating enzymes as oncogenes, non-oncogenes and tumor suppressors. Cell Cycle 8:1688–1697
Huth JR et al (2005) ALARM NMR: a rapid and robust experimental method to detect reactive false positives in biochemical screens. J Am Chem Soc 127:217–224. doi:10.1021/ja0455547
Jamal S, Scaria V (2013) Cheminformatic models based on machine learning for pyruvate kinase inhibitors of Leishmania mexicana. BMC Bioinformatics 14:329. doi:10.1186/1471-2105-14-329
Jamal S, Periwal V, Scaria V (2013) Predictive modeling of anti-malarial molecules inhibiting apicoplast formation. BMC Bioinformatics 14:55. doi:10.1186/1471-2105-14-55
Japkowicz N (2000) The class imbalance problem: significance and strategies. In: Proceedings of the International Conference on Artificial Intelligence
Jensen LJ, Saric J, Bork P (2006) Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Genet 7:119–129. doi:10.1038/nrg1768
Jensen BF, Vind C, Padkjaer SB, Brockhoff PB, Refsgaard HH (2007) In silico prediction of cytochrome P450 2D6 and 3A4 inhibition using Gaussian kernel weighted k-nearest neighbor and extended connectivity fingerprints, including structural fragment analysis of inhibitors versus noninhibitors. J Med Chem 50:501–511. doi:10.1021/jm060333s
Jones MJ, Colnaghi L, Huang TT (2012) Dysregulation of DNA polymerase kappa recruitment to replication forks results in genomic instability. EMBO J 31:908–918. doi:10.1038/emboj.2011.457
Kannouche PL, Wing J, Lehmann AR (2004) Interaction of human DNA polymerase eta with monoubiquitinated PCNA: a possible mechanism for the polymerase switch in response to DNA damage. Mol Cell 14:491–500
Kennedy RD, D'Andrea AD (2006) DNA repair pathways in clinical practice: lessons from pediatric cancer susceptibility syndromes. J Clin Oncol 24:3799–3808
Kim H, D'Andrea AD (2012) Regulation of DNA cross-link repair by the Fanconi anemia/BRCA pathway. Genes Dev 26:1393–1408. doi:10.1101/gad.195248.112
Kim JM, Parmar K, Huang M, Weinstock DM, Ruit CA, Kutok JL, D'Andrea AD (2009) Inactivation of murine Usp1 results in genomic instability and a Fanconi anemia phenotype. Dev Cell 16:314–320. doi:10.1016/j.devcel.2009.01.001
Liang Q et al (2014) A selective USP1-UAF1 inhibitor links deubiquitination to DNA damage responses. Nat Chem Biol 10:298–304. doi:10.1038/nchembio.1455
Liu K, Feng J, Young SS (2005) PowerMV: a software environment for molecular viewing, descriptor generation, data analysis and hit evaluation. J Chem Inf Model 45:515–522. doi:10.1021/ci049847v
Melville JL, Burke EK, Hirst JD (2009) Machine learning in virtual screening. Comb Chem High Throughput Screen 12:332–343
Mistry H et al (2013) Small-molecule inhibitors of USP1 target ID1 degradation in leukemic cells. Mol Cancer Ther 12:2651–2662. doi:10.1158/1535-7163.MCT-13-0103-T
Murai J, Yang K, Dejsuphong D, Hirota K, Takeda S, D'Andrea AD (2011) The USP1/UAF1 complex promotes double-strand break repair through homologous recombination. Mol Cell Biol 31:2462–2469. doi:10.1128/mcb.05058-11
Nijman SM, Huang TT, Dirac AM, Brummelkamp TR, Kerkhoven RM, D'Andrea AD, Bernards R (2005a) The deubiquitinating enzyme USP1 regulates the Fanconi anemia pathway. Mol Cell 17:331–339
Nijman SM, Luna-Vargas MP, Velds A, Brummelkamp TR, Dirac AM, Sixma TK, Bernards R (2005b) A genomic and functional inventory of deubiquitinating enzymes. Cell 123:773–786
Oestergaard VH et al (2007) Deubiquitination of FANCD2 is required for DNA crosslink repair. Mol Cell 28:798–809. doi:10.1016/j.molcel.2007.09.020
Periwal V, Rajappan JK, Jaleel AU, Scaria V (2011) Predictive models for anti-tubercular molecules using machine learning on high-throughput biological screening datasets. BMC Res Notes 4:504. doi:10.1186/1756-0500-4-504
Periwal V, Kishtapuram S, Scaria V (2012) Computational models for in vitro anti-tubercular activity of molecules based on high-throughput chemical biology screening datasets. BMC Pharmacol 12:1. doi:10.1186/1471-2210-12-1
Quinlan JR (1993) C4.5 programs for machine learning. Morgan Kaufmann Publishers, San Francisco
Sacco JJ, Coulson JM, Clague MJ, Urbe S (2010) Emerging roles of deubiquitinases in cancer-associated pathways. IUBMB Life 62:140–157. doi:10.1002/iub.300
Schierz AC (2009) Virtual screening of bioassay data. J Cheminform 1:21. doi:10.1186/1758-2946-1-21
Shen J, Cheng F, Xu Y, Li W, Tang Y (2010) Estimation of ADME properties with substructure pattern recognition. J Chem Inf Model 50:1034–1041. doi:10.1021/ci100104j
Sims AE et al (2007) FANCI is a second monoubiquitinated member of the Fanconi anemia pathway. Nat Struct Mol Biol 14:564–567. doi:10.1038/nsmb1252
Sud M (2010) MayaChemTools. http://www.mayachemtools.org/

Ulrich HD, Walden H (2010) Ubiquitin signalling in DNA replication and repair. Nat Rev Mol Cell Biol 11:479–489. doi:10.1038/nrm2921

Villamil MA, Chen J, Liang Q, Zhuang Z (2012) A noncanonical cysteine protease USP1 is activated through active site modulation by USP1-associated factor 1. Biochemistry 51:2829–2839. doi:10.1021/bi3000512

Wang Y et al (2012) PubChem's BioAssay Database. Nucleic Acids Res 40:D400–D412. doi:10.1093/nar/gkr1132

Williams SA et al (2011) USP1 deubiquitinates ID proteins to preserve a mesenchymal stem cell program in osteosarcoma. Cell 146:918–930. doi:10.1016/j.cell.2011.07.040

Zhuang Z, Johnson RE, Haracska L, Prakash L, Prakash S, Benkovic SJ (2008) Regulation of polymerase exchange between Poleta and Poldelta by monoubiquitination of PCNA and the movement of DNA polymerase holoenzyme. Proc Natl Acad Sci U S A 105:5361–5366. doi:10.1073/pnas.0801310105