



HHS Public Access

Author manuscript

J Comput Aided Mol Des. Author manuscript; available in PMC 2017 September 17.

Published in final edited form as:

J Comput Aided Mol Des. 2014 June ; 28(6): 631–646. doi:10.1007/s10822-014-9748-9.

Design, synthesis and experimental validation of novel potential chemopreventive agents using random forest and support vector machine binary classifiers

Brienne Sprague,

Department of Chemistry, Rutgers University, 315 Penn St., Camden, NJ 08102, USA

Qian Shi,

Natural Products Research Laboratories, UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, 310 Beard Hall, CB# 7568, Chapel Hill, NC 27599-7568, USA

Marlene T. Kim,

Department of Chemistry, Rutgers University, 315 Penn St., Camden, NJ 08102, USA

The Rutgers Center for Computational and Integrative Biology, Camden, NJ 08102, USA

Liyang Zhang,

Pfizer Worldwide Research and Development, Groton, CT 06340, USA

Alexander Sedykh,

Multicase Inc, 23811 Chagrin Blvd, Suite 305, Beachwood, OH 44122, USA

Eiichiro Ichiishi,

Department of Internal Medicine, International University of Health and Welfare Hospital, Tochigi 329-2763, Japan

Harukuni Tokuda,

Department of Complementary and Alternative Medicine Clinical R&D, Kanazawa University of Graduate School of Medical Science, Kanazawa 920-8640, Japan

Kuo-Hsiung Lee, and

Natural Products Research Laboratories, UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, 310 Beard Hall, CB# 7568, Chapel Hill, NC 27599-7568, USA

Chinese Medicine Research and Development Center, China Medical University and Hospital, Taichung, Taiwan

Hao Zhu

Department of Chemistry, Rutgers University, 315 Penn St., Camden, NJ 08102, USA

The Rutgers Center for Computational and Integrative Biology, Camden, NJ 08102, USA

Abstract

Correspondence to: Qian Shi; Hao Zhu.

Electronic supplementary material The online version of this article (doi:10.1007/s10822-014-9748-9) contains supplementary material, which is available to authorized users.

Compared to the current knowledge on cancer chemotherapeutic agents, only limited information is available on the ability of organic compounds, such as drugs and/or natural products, to prevent or delay the onset of cancer. In order to evaluate chemical chemopreventive potentials and design novel chemopreventive agents with low to no toxicity, we developed predictive computational models for chemopreventive agents in this study. First, we curated a database containing over 400 organic compounds with known chemoprevention activities. Based on this database, various random forest and support vector machine binary classifiers were developed. All of the resulting models were validated by cross validation procedures. Then, the validated models were applied to virtually screen a chemical library containing around 23,000 natural products and derivatives. We selected a list of 148 novel chemopreventive compounds based on the consensus prediction of all validated models. We further analyzed the predicted active compounds by their ease of organic synthesis. Finally, 18 compounds were synthesized and experimentally validated for their chemopreventive activity. The experimental validation results paralleled the cross validation results, demonstrating the utility of the developed models. The predictive models developed in this study can be applied to virtually screen other chemical libraries to identify novel lead compounds for the chemo-prevention of cancers.

Keywords

Cheminformatics; Chemoprevention; Natural products; QSAR; Virtual screening

Introduction

Cancer is listed among the major causes of mortality in the world. In 2013, cancer was ranked as the second leading cause of death in the United States. A recent report from the American Cancer Society revealed that, statistically, the lifetime chances of developing cancers are as high as 1 in 3 for women and 1 in 2 for men [1]. Current strategies in cancer patient treatments, such as chemotherapy, have met with clinical success. However, most chemotherapeutic agents have severe side effects, which negatively impact a cancer patient's quality of life [2]. For this reason, chemoprevention, which normally uses either natural or synthetic compounds with low toxicity, was employed to impede, halt, or reverse the carcinogenesis process before a tumor can develop [3], especially for patients at high risk. The different stages of chemoprevention research have been extensively reviewed [4-8].

As a relatively new area of cancer research, there is a high demand for efficient methods to identify novel chemoprevention agents. To date, a standardized method for identifying chemopreventive compounds has not been developed. But several in vitro methods have been used by different research groups to evaluate potential chemopreventive agents [9-11]. Most of these methods measure chemopreventive activity on the basis of similar general principles. These approaches involve measuring cellular expression of a protein in a human cancer cell line with and without the test compound. A positive impact is noted by a decrease in expression, which indicates interference with a potential cancer-inducing pathway [12-18]. The Epstein-Barr virus early antigen (EBV-EA) activation assay is one such test recognized as a primary screening method for assessing antitumor promoting properties, through inhibition of Protein Kinase C (PKC) activity, which serves as a major

receptor for 12-*O*-tetradecanoylphorbol-13-acetate (TPA) [19, 20]. In this test, the experiment is carried out in Raji cells, a human lymphoma cell line with an Epstein–Barr virus genome. The assay is less tumor type specific, which limits its applicability in mechanistic studies of chemoprevention; however, it can measure the chemopreventive effects of a particular agent on the promotion and progression phases of carcinogenesis, and usually results in parallel outcomes to in vivo animal models. It is considered to be a useful in vitro assay to initially screen chemopreventive agents [12, 13, 18].

Currently, only a few studies have employed computational modeling in the area of chemoprevention. Among them, several Quantitative structure–activity relationship (QSAR) studies used simple modeling approaches, such as Multiple linear regression (MLR) or Partial least square (PLS), to model a limited number of chemopreventive agents. For example, Bertosa et al. [21] used PLS to develop QSAR models for 59 amides and quinolones. The antitumor activities of these compounds were tested against MiaPaCa-2 (pancreatic carcinoma) and MCF-7 (breast carcinoma) cells. In another study, Saeed et al. [22] used linear regression to generate models for six curcumin derivatives. In a recent published work, Aleksic et al. [23] reported a three-dimensional (3D) QSAR study of substituted heterocyclic quinolones. Nineteen compounds with similar quinolone scaffolds were synthesized and tested for antitumor activity against multiple cell lines in this study. The results were modeled with commercial modeling software. Other receptor-based modeling studies, which used molecular docking analysis, have focused on using known cancer tumor targets and their interaction with specific chemopreventive compounds. For example, a recent study tested the binding affinities of curcumin derivatives against several well-known cancer targets [24]. The results revealed a favorable correlation between the two test compounds and their binding affinity for cancer targets, thus, implying a functional role as enzyme inhibition activators. Although previous modeling studies achieved certain successes (e.g., molecular docking studies usually helped to explain the binding mechanisms of the chemopreventive agents to the relevant receptors), they have restricted predictive ability due to the limited number of compounds used in the studies.

In this study, we proposed to develop QSAR models based on a large set of chemopreventive agents tested by the EBV-EA assay and to apply the resulting models to design new lead agents. To this end, we curated a chemopreventive database by collecting compounds tested by the EBV-EA assay from published research articles [25-51]. We used an in-house tool to automatically generate an activity endpoint based on the original multi-dose chemopreventive response data. Various QSAR models were developed and validated. Then, these models were used to virtually screen over 23,000 natural products and their derivatives to identify novel chemopreventive agents. Finally, 18 lead compounds resulting from the prediction results were synthesized and experimentally validated using the same EBV-EA assay.

Methods

Chemopreventive agent data set

The chemopreventive agent database was generated by collecting data from different journal papers published during the past decade [25-51]. All compounds in the database were tested

using the EBV-EA assay by the same standard methodology. The original database contained compounds that were reported in different papers. After removing duplicated compounds by harmonizing their activities, we had 405 unique compounds left in the final curated chemopreventive agent database.

The EBV-EA assay data were treated by the CurveP algorithm to ensure monotonicity of each dose–response curve and to convert it into a numeric value, LogCurveP (a log₁₀-transformed fingerprint), which was used as a numeric indicator of activity. This algorithm was developed in our laboratory in a prior study [52]. Briefly, the response at each of the various test concentrations was represented by two bits (00, 01, 10, 11) for coding four categories (0, 25, 50, and 75 % + relative inhibition). Then, these bits were concatenated from lowest to highest test concentrations and resulted in an eight bit value (CurveP). If nonzero, the CurveP was then log₁₀-transformed. The chosen activity threshold of 1.25 corresponds to strong inhibition at the two highest test concentrations. Based on these factors, the chemopreventive agent database contained 204 “actives” categorized as class-1 and 201 “marginal actives” categorized as class-2. All 405 compounds and their chemopreventive activity (both the original data and the LogCurveP results) are listed in a supplemental file (Supplemental Table 1).

EBV-EA assay

Raji cells (10⁶ cells/mL) were incubated at 37 °C for 48 h in RPMI-1640 medium with 10 % Fetal calf serum (FCS), *n*-butyric acid (4 mmol), TPA (32 pmol), and test compounds. Smears were made from the cell suspension, and EBV-EA inducing cells were stained by an indirect immunofluorescence technique. In each assay, at least 500 cells were counted and the number of stained cells (positive cells) was recorded. The EBV-EA-inhibiting activity of the test compound was estimated on the basis of the percentage of the number of positive cells compared with that of the control without the test compound. Cell viability was assayed by the Trypan Blue staining method. For the determination of cytotoxicity, the cell viability was required to be more than 60 % [53]. Each compound was measured at four concentrations of 0.32, 3.2, 16, and 32 nmol, representing 10-, 100-, 500-, 1,000-fold of TPA (32 pmol). Each measurement was repeated three times for each test compound concentration and the average values of the three readout data were used.

Chemical descriptors

The chemical descriptors used in this study were obtained from Dragon version 6.0 (Talet SRL, Milano, Italy) and Molecular Operating Environment (MOE) version 2011. The Dragon descriptors include E-state values and E-state counts, constitutional descriptors, topological descriptors, walk and path counts, connectivity and information indices, 2D autocorrelations, Burden eigenvalues, molecular properties, Kappa, hydrogen bond acceptor/donor counts, molecular distance edge, and molecular fragment counts. The MOE descriptors include topological indices, structural keys, E-state indices, physical properties (i.e., LogP, molecular weight, and molar refractivity), and topological polar surface area. Over 4,000 Dragon descriptors were initially generated, but most of them were redundant. We removed redundant Dragon descriptors by using pairwise comparisons between each of the two descriptor pairs. If the correlation between two descriptors of our 405 compounds

was high (correlation coefficient >0.95), one of them was randomly selected and removed. Eventually, 688 Dragon descriptors were left for this study. MOE generated 186 descriptors, all of which were used in the modeling process.

Modeling approaches

This study used the Random forest (RF) and Support vector machine (SVM) algorithms available in R.2.15.1. These two algorithms have been employed in several of our previous modeling studies for various biological activities [54, 55].

The entire Combinatorial QSAR (Combi-QSAR) modeling workflow is shown in Fig. 1. Individual models were developed using a combination of Dragon or MOE descriptors and RF or SVM algorithms. This technique resulted in four different models: Dragon-RF, Dragon-SVM, MOE-RF, and MOE-SVM. The results for each classification model were averaged to generate consensus predictions, which will be further referred to as a consensus model.

All models were validated using five-fold external cross validation. Briefly, the original chemopreventive dataset was randomly divided into five equal subsets. One subset was used as the validation set (20 % of the original set) and the other four subsets (80 % of the original set) were used as the training sets. The training sets were used to develop the models and the resulting models were used to predict the left-out validation set. This procedure was repeated five times, so that each compound was used for validation purposes once.

Robustness of QSAR models was verified using a Y-randomization (randomization of response) approach as described by Tropsha and coworkers [56-58]. We randomly assigned the activities of the modeling set compounds into class-1 or -2. Then, we developed QSAR models using the same protocol as for compounds with actual experimental results. The purpose of this procedure was to see if statistically significant QSAR models could be obtained for the original data, but could not be developed with randomized activities. The Y-randomization tests for each combination of modeling approach and descriptor were repeated five times.

Universal criteria for model evaluation

Because various modeling approaches and different descriptors were used in the modeling process, universal statistical metrics were needed to evaluate the performance of the models developed individually. The results were harmonized by using sensitivity (percentage of class-1 compounds predicted correctly), specificity (percentage of class-2 compounds predicted correctly), and CCR (correct classification rate or balanced accuracy). These parameters are defined as follows:

$$\% \text{ sensitivity} = \left(\frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \right) \times 100 \quad (1)$$

$$\% \text{ specificity} = \left(\frac{\text{true negatives}}{\text{true negatives} + \text{false positives}} \right) \times 100 \quad (2)$$

$$\% \text{ CCR} = \left(\frac{\text{sensitivity} + \text{specificity}}{2} \right) \times 100 \quad (3)$$

Synthesis of selected compounds

Eighteen compounds derived from the predicted results were designated for chemical synthesis. Among them, twelve compounds (**1–12**) resulted from the class-1 compound list and six compounds (**13–18**) were derived from the class-2 prediction set (Table 2). Compound selection from more than 100 predicted leads was based on the chemical capability of producing the molecules through basic organic synthesis and the availability of chemical reagents for making the compound. Compound **12** was purchased from Aldrich. All final compounds were structurally confirmed by mass spectrometry (Shimadzu LCMS-2010 ESI-MS) and proton nuclear magnetic resonance spectroscopy (^1H NMR) [Varian 400 MHz with tetramethylsilane (TMS) as the internal standard]. Melting points were determined on a Fisher-John melting point apparatus and are uncorrected. CombiFlash[®] chromatographic system (Isco Companion) with a Grace silica gel cartridge was used for general separation and purification. Preparative thin layer chromatography (PTLC) on silica gel plates (Kieselgel 60, F254, 1.50 mm) was also used for separation and purification. Precoated silica gel plates (Kieselgel 60, F254, 0.25 mm) were used for Thin layer chromatography (TLC) analysis. All reagents and solvents were purchased from Aldrich, Fisher, VWR, and other vendors. Some chemicals were used after purification, and others were used as purchased.

Results and discussions

The overview of our chemopreventive agent database

We analyzed the structural similarities between the compounds in the dataset by performing a Principal component analysis (PCA) on the chemical descriptors. After generating the principal components using the 186 MOE descriptors for all of the compounds in the database, we selected the top three most important components to create a 3D plot (Fig. 2) for all 405 compounds. Considering the 186 MOE descriptors that we used, these three principal components captured around 30 % of the variance in our database. In this way, we could visualize the chemical similarity between modeling set compounds in this 3D plot (Fig. 2). According to this analysis, not surprisingly, many compounds were chemically similar, since they are derivatives of several known chemopreventive agents (e.g., curcumin). But there were several structural outliers that were dissimilar to the majority of the compounds. Some previous studies showed that removing structural outliers before the modeling process was beneficial to the results of the QSAR models [55, 59, 60]. However, in our study, we kept these outliers, since they were only a small portion (~1 %) of the whole

dataset. Furthermore, removing the outliers did not improve the resulting models (data not shown).

Chemopreventive activity endpoint

All 405 compounds were tested in the EBV-EA assay at four different doses (10, 100, 500 and 1,000 times the TPA dose) and the chemopreventive activity reported as relative percentage induction of TPA-mediated EBV-EA activation (Fig. 3a). Most of the compounds showed significant activity at high dose levels, but half of the compounds exhibited no activity at the lowest dose. We next applied a method previously developed in our laboratory and successfully applied in prior studies in which the multi-dose response data for each compound was converted into a meaningful endpoint that could be used for modeling purposes [52]. Figure 3b shows the transformed Log-CurveP results based on the original four dose induction response data from Fig. 3a. Noticeably, a clear threshold (LogCurveP = 1.25) was present (see middle of Fig. 3b), which could be used to distinguish “actives” and “marginal actives”. It should be emphasized that the definition of these two categories is somewhat arbitrary, since most of the compounds showed significant activity in the high dose testing. However, this strategy gave us a criterion to differentiate the compounds that are likely to have high efficacy from the remaining compounds. On this basis, the actual chemopreventive agent modeling database contained 204 “actives” and 201 “marginal actives”.

Modeling results

We developed four individual and one consensus model for the 405 compounds (204 actives and 201 marginal actives). The fivefold external cross validation results for all of the models are shown in Table 1. The sensitivity, specificity, and CCR metrics for the four individual models ranged from 57 to 75, 61 to 74, and 59 to 74 %, respectively. The SVM-MOE model had the lowest predictivity (CCR = 59 %), and the RF-DRG model had the highest predictivity (CCR = 74 %). The consensus model showed equivalent statistics with sensitivity, specificity, and CCR all equal to 69 %.

Y-randomization tests were also performed for the modeling set. After five time random assignments of class-1 or -2 to the 405 compounds, we developed four individual QSAR models. The average CCR values obtained from five-fold cross validation for all four individual models with randomized classes were around 0.5, indicating that randomization of the classifications did not result in meaningful models. In addition, we used Pearson's Chi squared test to calculate the χ^2 and p values for the prediction results obtained using actual and randomized classes [61]. The improvement achieved by our real QSAR models, compared with those obtained by randomized classes, was statistically significant ($\chi^2 > 30$ and $p < 0.0001$).

Figure 4 shows the Receiver Operating Characteristic (ROC) of all four individual models. The Area under the curve (AUC) is another metric to evaluate the performance of each model. The RF models (AUC = 0.83 and 0.80), either with Dragon and MOE descriptors, were superior to the SVM models (AUC = 0.72 and 0.68).

Furthermore, we applied Consensus prediction thresholds (CPT), as cited in one of our previous studies [62], to the prediction results. Since all of the prediction values from the individual models ranged from 0 to 1, we initially used the 0.5 value as a single threshold to distinguish compounds predicted as class-1 (CPT ≥ 0.5) or class-2 (CPT < 0.5). However, as shown in Fig. 4, the use of more restrictive thresholds improved the predictivities of all models. Consequently, compounds with CPT values around 0.5 should be considered “inconclusive”. Based on the results in Fig. 4, we removed these inconclusive predictions by using two arbitrary, but reasonable, CPT thresholds to classify compounds as actives (CPT > 0.7) and marginal actives (CPT < 0.3).

The application of CPT to define the prediction results together with the removal of “inconclusive” compounds clearly enhanced the predictivity of all models, including the consensus model (Table 1). For example, the sensitivity, specificity, and CCR metrics of the consensus model increased to 83, 82, and 82 %, respectively (Table 1). However, the tradeoff was to decrease the coverage of this model from 100 to 46 %. In addition, the coverage of the individual models ranged from 45 to 58 % after applying CPT and excluding inconclusive compounds. Since we expect that the models developed in this project will be used to screen large chemical libraries and prioritize a small portion of “hits” for experimental validation, we feel that it is reasonable to sacrifice prediction coverage to increase predictivity.

Virtual screening

Once we developed and validated our predictive QSAR models, they could then be used to screen new compounds for chemopreventive activity. Since chemopreventive agents usually must be administered for a long period of time, low toxicity and fewer side effects are essential factors in the design of new agents. Therefore, we used the ZINC natural derivatives (ZND) library that contains over 23,000 natural product molecules and their derivatives for screening purposes [63]. The original ZND database was curated to remove duplicates, including compounds that overlapped with our existing dataset as well as compounds that could not be handled by our program. This process resulted in a total of over 23,385 unique compounds available for virtual screening. Next, we evaluated these compounds with all four individual models to prioritize and choose hits. We prioritized those compounds that were calculated as “active” and excluded those compounds that were predicted as “inconclusive” (prediction values between 0.3 and 0.7) based on the consensus predictions of all four individual models. As another selection criterion, we produced a combined score by summation of all individual model prediction values. By applying both selection criteria, we ultimately selected 148 compounds from the ZND library. These compounds were predicted to be active hits by all four models (individual prediction values were all above 0.7), as well as had the highest combined scores based on summations of all four predictions. For comparison purposes, we intentionally selected 45 compounds that were predicted to be “marginal active” by all models using the same strategy as for the active hits.

Chemical synthesis and experimental validation by in vitro EBV-EA inhibition

From the 148 class-1 structures and 45 class-2 structures predicted by virtual screening, we selected 12 compounds from the active lead set and 6 compounds from the marginal active set. Our selection rationale included chemical synthesis capability and availability, as well as the SAR (structure–activity relationship) profile from previously reported literature on chemopreventive agents. For instance, favored structural groups in most of the active predictions from our virtual screening included phenolic groups and a biphenyl moiety with a conjugated carbonyl system and/or a 4*H*-chromen-4-one. These chemical structures are common features of some known chemopreventive agents, such as curcumin derivatives and flavonoids [64–66]. Selected “active” compounds **1–4** are structural mimics of curcuminoids, while compounds **6** and **8–10** belong to the flavonoid chemical class.

We next synthesized the 18 chosen compounds. Compounds **1**, **3**, and **13–16** were prepared by reaction of an appropriately substituted benzoic acid (compound **16**) or cinnamic acid (compounds **1**, **3**, and **13–15**) with an appropriate amine in the presence of the coupling reagent EDCI hydrochloride and the catalyst DAMP (Scheme 1). Compounds **2** and **4** were obtained by subsequent demethylation of **1** and **3** with BBr₃ at low temperature. Compound **5** was synthesized by reaction of 1-(2-hydroxy-5-methoxyphenyl)ethanone with 2,3-dimethoxybenzaldehyde. Treatment of **5** with sodium acetate in aqueous ethanol and heating to reflux gave cyclized compound **5a**, which underwent demethylation with BBr₃ in methylene chloride yielding compound **6** (Scheme 2). Unexpectedly, the ring-opened product **7** was also obtained during the demethylation process, probably due to the instability of the 2*H*-pyran-4(3*H*)-one moiety under the reaction conditions (Scheme 2). Compound **17** was prepared by reaction of naphthalen-1-ol with methyl 2-chloroacetate in the presence of potassium carbonate. Compound **18** was afforded by demethylation of **17** with trimethylstannanol (Scheme 3). Compounds **8–11** were synthesized by heating 1-(2-hydroxy-5-methoxyphenyl)ethanone and 3,4-dimethoxybenzoyl chloride in pyridine (Scheme 4). After treatment of the resulting compound with potassium hydroxide followed by acidification, the cyclized compound **8** was obtained. Selective demethylation of compound **8** afforded compounds **9–11**.

All 18 synthesized compounds were evaluated for chemoprevention activities measured as inhibition of TPA-induced EBV-EA expression in Raji cells. These 18 compounds and their relevant response data are shown in Table 2. The value corresponding to each compound indicates a relative ratio to the positive control TPA on activation of EBV-EA expression in Raji cells. Unsurprisingly, compounds derived from the predicted class-1 set (compounds **1–12**), especially compounds **8–12** with a flavonoid structural scaffold, generally showed more potent inhibition of EBV-EA expression than those derived from the class-2 set (**13–18**). Based on our definition of class-1 and -2 as mentioned above, all predicted “marginal active” compounds (compounds **13–18**) were experimentally proved to be correctly predicted. Among all actives (compounds **1–12**), compounds **5–12** were True positives (TP); however, compounds **1–4** were False positives (FP). Structurally, compounds **1–4** are derived from curcuminoids. However, in comparison with curcumin, compounds **1–4** were weaker inhibitors in the validation assay, especially at the higher concentration levels (Table 2). Thus, the high activity of curcumin derivatives in our modeling set was the major reason for

these FP predictions. This result also provided us with important information on revising/optimizing curcumin derivatives as chemoprevention agents. In summary, the experimental validation showed 67 % sensitivity, 100 % specificity, and 83 % CCR. Although the number of experimentally tested compounds was not great, the experimentally validated results clearly demonstrate that the newly developed cheminformatics models can be used to screen new chemical libraries and prioritize novel hits for future development.

Further Structure–Activity Relationship profiles resulting from this study indicated that phenolic substitution in the molecule enhanced the EBV-EA inhibition ability. Flavonoids **9–12** with multiple phenolic hydroxyl groups displayed 100 % inhibition at the highest tested concentration and 24–28 % inhibition even at 1×10^2 mol ratio to TPA (32 pmol). Compound **9** was the most potent analog among the tested compounds showing significant inhibition even at concentrations as low as 1×10 mol ratio to TPA. Interestingly, compound **7**, a ring-opened analog bearing four phenolic hydroxy groups, was less potent than its ring-closed analog **9** against EBV-EA activation. These results suggested that the flavonoid skeleton is essential for the inhibition activity and phenolic groups enhance the inhibition potency.

To evaluate the novelty of the 18 new compounds, we analyzed the major chemical features of the compounds and compared them to those existing in the modeling set. To this end, chemical scaffolds were generated and compared for the 18 new compounds against the 408-compound dataset. All compounds were reduced to core fragments (or “scaffolds”) based on the method reported in a previous study [67]. Eight unique scaffolds were generated out of the 18 compounds (Fig. 5), and 167 unique scaffolds out of the 408 compounds. By comparison, 50 % of the prior scaffolds (4 out of 8) were novel and did not exist in the 408-compound dataset (Fig. 5).

In summary, we employed a Combi-QSAR workflow to develop predictive models for a database consisting of 405 chemopreventive agents, which were all tested by EBV-EA assay. We used our in-house tool to define a chemoprevention activity endpoint that was suitable for modeling purposes. The resulting four individual models were validated by a five-fold cross validation procedure. The consensus prediction showed superior performance compared with that of the individual models. For this reason, we used all four individual models to virtually screen a large chemical library. Eighteen “hits”, 12 from a class-1 “active” set and 6 from a class-2 “marginal active” set, were finally selected, synthesized, and then validated by the same EBV-EA assay. The validation results indicated that the compounds derived from the “active” prediction were more potent EBV-EA activation inhibitors than the compounds derived from the “marginal active” prediction. Both the cross validation and experimental validation results showed that our developed models are suitable for designing novel chemopreventive agents by prioritizing novel molecules for experimental testing and further development.

Experiments

Chemical synthesis

General synthesis procedures for compounds 1–4 and 13–16—Substituted cinnamic acid (1 eq., for compounds **1,3, 13–15**,) or benzoic acid (1 eq., for compound **16**) was dissolved in DMF. EDCI hydrochloride (1.5 eq.) and 10 % (mol ratio) of DMAP were added. After being stirred at r.t. for 30 min. an appropriate amine (1.5 eq.) was added. The resulting mixture was stirred at r.t. overnight. The solid was removed by filtration and the filtrate was concentrated under vacuum. The residue was partitioned in EtOAc and water, and the organic portion was washed twice with water. After drying over Na₂SO₄, the crude product was purified by column chromatograph through a combiflash system with hexanes/EtOAc as eluent.

(E)-3-(3-hydroxyphenyl)-N-(3-methoxybenzyl)-N-methylacrylamide (1)—White crystalline solid. ¹H NMR (400 MHz, CDCl₃): δ 7.75 (d, *J* = 15.6 Hz, 1H), 7.29–7.02 (m, aromatic H, 4H), 6.92–6.73 (m, 4H), 6.55 (d, *J* = 17.2 Hz, 1H), 4.66 (d, *J* = 16.0 Hz, PhCH₂N–, 2H), 3.82 (s, OCH₃, 3H), 3.07 (s, NCH₃, 3H); ESI MS *m/z* 298.20 (M + H)⁺.

(E)-N-(3-hydroxybenzyl)-3-(3-hydroxyphenyl)-N-methylacrylamide (2)—White crystalline solid. ¹H NMR (400 MHz, CDCl₃): δ 7.65 (d, *J* = 15.6 Hz, 1H), 7.24–6.69 (m, aromatic H, 7H), 6.02 (d, *J* = 17.2 Hz, 1H), 5.82 (d, *J* = 17.2 Hz, 1H), 4.62 (d, *J* = 13.6 Hz, PhCH₂N–, 2H), 3.02 (s, NCH₃, 3H); ESI MS *m/z* 283.91 (M + H)⁺.

(E)-N-(3-methoxybenzyl)-3-(3-methoxyphenyl)acrylamide (3)—White crystalline solid. ¹H NMR (400 MHz, CDCl₃): δ 7.61 (d, *J* = 15.6 Hz, 1H), 7.27–7.22 (m, aromatic-H, 2H), 7.07 (d, *J* = 7.6 Hz, 1H), 6.99 (s, aromatic-H, 1H), 6.99–6.79 (m, aromatic-H, 4H), 6.38 (d, *J* = 15.6 Hz, 1H), 5.96 (s, br, –NH–, 1H), 4.52 (d, *J* = 6.0 Hz, PhCH₂N–, 2H), 3.79, 3.78 (s, OCH₃ × 2, 3H each); ESI MS *m/z* 298.20 (M + H)⁺.

(E)-N-(3-hydroxybenzyl)-3-(3-hydroxyphenyl)acrylamide (4)—Off-white crystalline solid; ¹H NMR (400 MHz, CDCl₃): δ 7.47 (d, *J* = 15.6 Hz, 1H), 7.19–7.10 (m, aromatic-H, 2H), 7.01 (d, *J* = 7.2 Hz, 1H), 6.95 (s, aromatic-H, 1H), 6.78–6.73 (m, aromatic-H, 3H), 6.66–6.64 (m, aromatic-H, 1H), 6.57 (d, *J* = 15.6 Hz, 1H), 4.40 (s, PhCH₂N–, 2H); ESI MS *m/z* 270.21 (M + H)⁺.

(E)-3-(benzo[d][1,3]dioxol-5-yl)-N-(3,4-dimethoxyphenethyl)acrylamide (13)—Off-white crystalline solid. ¹H NMR (400 MHz, CDCl₃): δ 7.50 (d, *J* = 15.2 Hz, 1H), 6.95 (s, aromatic H, 1H), 6.93 (d, *J* = 1.6 Hz, aromatic H, 1H), 6.80–6.71 (m, aromatic H, 4H), 6.11 (d, *J* = 15.2 Hz, 1H), 5.96 (s, methylene H, 2H), 5.52 (br. NH, 1H), 3.84 (s, OCH₃, 6H), 3.60 (q, NHCH₂–, 2H), 2.78 (t, *J* = 7.2 Hz, 2H); ESI MS *m/z* 356.20 (M + H)⁺.

(E)-3-(benzo[d][1,3]dioxol-5-yl)-N-(4-chlorophenethyl)acrylamide (14)—White crystalline solid. ¹H NMR (400 MHz, CDCl₃): δ 7.50 (d, *J* = 15.2 Hz, 1H), 7.25 (d, *J* = 8.4 Hz, aromatic H, 2H), 7.12 (d, *J* = 8.4 Hz, aromatic H, 2H), 6.95 (s, aromatic H, 1H), 6.93 (d, *J* = 2.0 Hz, aromatic H, 1H), 6.76 (dd, *J* = 1.2, 7.6 Hz, aromatic H, 1H), 6.11 (d, *J* = 15.2 Hz,

1H), 5.95 (s, methylene H, 2H), 5.52 (br. NH, 1H), 3.60 (q, $J = 6.8$ Hz, NHCH_2 -, 2H), 2.83 (t, $J = 7.2$ Hz, 2H); ESI MS m/z 330.15 (M + H)⁺.

(E)-3-(benzo[d][1,3]dioxol-5-yl)-N-(3-phenylpropyl)acrylamide (15)—White crystalline solid. ¹H NMR (400 MHz, CDCl_3): δ 7.47 (d, $J = 15.6$ Hz, 1H), 7.28–7.23 (m, aromatic H, 2H), 7.18–7.16 (m, aromatic H, 3H), 6.95 (d, $J = 0.8$ Hz, aromatic H, 1H), 6.93 (d, $J = 1.2$ Hz, aromatic H, 1H), 6.76 (d, $J = 8.4$ Hz, aromatic H, 1H), 6.11 (d, $J = 15.6$ Hz, 1H), 5.96 (s, methylene H, 2H), 5.52 (br. NH, 1H), 3.39 (q, $J = 6.8$ Hz, NHCH_2 -, 2H), 2.70 (t, $J = 7.6$ Hz, 2H), 1.88 (pent, $J = 7.2$ Hz, 2H); ESI MS m/z 310.21 (M + H)⁺.

(E)-3-(benzo[d][1,3]dioxol-5-yl)-N-(3-(dimethylamino)propyl)acrylamide (16)—Light yellow solid. ¹H NMR (400 MHz, CDCl_3): δ 8.83 (br. NH, 1H), 6.93 (dd, $J = 2.0$, 9.2 Hz, aromatic H, 1H), 6.88 (s, aromatic H, 1H), 6.81 (dd, $J = 2.0$, 9.2 Hz, aromatic H, 1H), 5.99 (s, methylene H, 2H), 3.33 (q, $J = 6.8$ Hz, NHCH_2 -, 2H), 2.42 (t, $J = 7.6$ Hz, 2H), 2.27 (s, $\text{N}(\text{CH}_3)_2$, 6H), 1.75 (pent, $J = 7.2$ Hz, 2H); ESI MS m/z 251.24 (M + H)⁺.

Syntheses of compounds 5–7—To a solution of 1-(2-hydroxy-5-methoxyphenyl)ethanone (3 mmol) in 10 mL of EtOH was added 2,3-dimethoxybenzaldehyde (1.05 eq.). 20 % KOH (5 mL aq.) was added dropwise. The resulting red mixture was stirred at r.t for 5 h with TLC monitoring, and then was poured into ice water, acidified with 2 N HCl to pH 2, and extracted with EtOAc. After purification through a combiflash column chromatography system, compound 5 was obtained. Compound 5 (0.6 mmol) was dissolved in 5 mL of EtOH and NaOAc (10 eq.) was added. The mixture was heated to reflux until the reaction was complete (approximately 24 h). The reaction mixture was poured into ice water and extracted with EtOAc. The crude product was purified by column chromatography on a combiflash system with hexanes/EtOAc as eluent to yield 5a. Compound 5a (0.545 mmol) was dissolved in 20 mL of anhydrous methylene chloride (CH_2Cl_2) and cooled to -78 °C. BBr_3 (1 M in CH_2Cl_2 , 4.5 eq.) was added slowly. The resulting mixture was stirred at -78 °C for 10 min, 0 °C for 10 min, and r.t. for 3 h with TLC monitoring. The reaction mixture cooled in an ice-bath then pured into ice water with stirring. The mixture was extracted with ethyl ether (Et_2O) three times and dried over Na_2SO_4 . The desired products **5** and **6** were obtained after purification over a combiflash system with $\text{CH}_2\text{Cl}_2/\text{MeOH}$ as eluent.

(E)-3-(2,3-dimethoxyphenyl)-1-(2-hydroxy-5-methoxyphenyl)prop-2-en-1-one (5)—Yellow crystalline solid. ¹H NMR (400 MHz, CDCl_3): δ 8.15 (d, $J = 16.0$ Hz, 1H), 7.68 (d, $J = 16.0$ Hz, 1H), 7.33 (d, $J = 2.8$ Hz, aromatic H, 1H), 7.24 (d, $J = 8.0$ Hz, aromatic H, 1H), 7.13–7.07 (m, aromatic H, 2H), 6.98–6.94 (m, aromatic H, 2H), 3.89 (s, OCH_3 , 3H), 3.88 (s, OCH_3 , 3H), 3.80 (s, OCH_3 , 3H); ESI MS m/z 315.21 (M + H)⁺.

2-(2,3-dihydroxyphenyl)-6-methoxychroman-4-one (6)—Light yellow solid. ¹H NMR (400 MHz, CDCl_3): δ 7.30 (d, $J = 3.2$ Hz, aromatic H, 1H), 7.14 (dd, $J = 3.2$, 8.0 Hz, aromatic H, 1H), 6.99 (d, $J = 8.0$ Hz, aromatic H, 1H), 6.97 (d, $J = 5.6$ Hz, aromatic H, 1H), 6.74 (td, $J = 2.0$, 8.0 Hz, aromatic H, 1H), 6.70 (t, $J = 8.0$ Hz, aromatic H, 1H), 5.73 (dd, $J = 3.2$, 12.0 Hz, 1H), 3.77 (s, OCH_3 , 3H), 2.97 (dd, $J = 12.8$, 30.0 Hz, 1H), 2.84 (dd, $J = 3.2$, 20.0 Hz, 1H); ESI MS m/z 285.21 (M + H)⁺.

(E)-3-(2,3-dihydroxyphenyl)-1-(2,5-dihydroxyphenyl)prop-2-en-1-one (7)—

Yellow crystalline solid. ¹H NMR (400 MHz, CDCl₃): δ 8.14 (d, *J* = 15.6 Hz, 1H), 7.84 (d, *J* = 15.6 Hz, 1H), 7.36 (d, *J* = 2.8 Hz, aromatic H, 1H), 7.11 (dd, *J* = 1.2, 8.0 Hz, aromatic H, 1H), 6.99 (dd, *J* = 2.8, 8.0 Hz, aromatic H, 1H), 6.83 (dd, *J* = 1.6, 8.0 Hz, aromatic H, 1H), 6.79 (d, *J* = 8.8 Hz, aromatic H, 1H), 6.70 (t, *J* = 8.0 Hz, aromatic H, 1H); ESI MS *m/z* 273.21 (M + H)⁺.

Syntheses of compounds 8–10—To a solution of 1-(2-hydroxy-5-methoxyphenyl)ethanone (1.84 mmol) in pyridine (3 mL) was added 3,4-dimethoxybenzoyl chloride (3 eq.). The resulting mixture was heated to reflux for 1 h. After cooling to r.t., the reaction mixture was pured into ice water with stirring, and solids started to precipitate. After storage in a refrigerator overnight, the off-white solid product (2-acetyl-4-methoxyphenyl 3,4-dimethoxybenzoate) was collected by filtration and dried in vacuo (0.4 g). The resulting compound (0.9 mmol) was further dissolved in 1 mL of pyridine, and 86 mg of KOH powder was added with stirring. The mixture was then heated at 50 °C for 1 h, then poured into 10 % H₂SO₄ (8 mL). A light brown solid precipitated. The solid was filtered and dissolved in 1.5 mL of EtOH containing 0.1 mL of H₂SO₄. The solution was heated to reflux for 1 h followed by alkalization to pH 10 with 20 % NaOH and refluxed for another 15 min. After cooling, the solution was neutralized with 10 % H₂SO₄ to give a solid, which was recrystallized from MeOH to afford compound 8 (0.121 g) as a dark gray crystalline solid. Compound 8 (0.096 mmol) was dissolved in 5 mL of anhydrous CH₂Cl₂ and cooled to –78 °C. Then, BBr₃ (1 M in CH₂Cl₂, 4.5 eq.) was added slowly. The resulting mixture was stirred at –78 °C for 10 min, 0 °C for 10 min, and r.t. for 2 h with TLC monitoring. The reaction mixture then was cooled in an ice-bath and poured into ice water. After stirring for 0.5 h, the mixture was extracted with Et₂O and dried over Na₂SO₄. The desired products **9–11** were obtained after column chromatography over a combiflash system with CH₂Cl₂/MeOH as eluent.

2-(3,4-dimethoxyphenyl)-6-methoxy-4H-chromen-4-one (8)—Yellow–brown crystalline solid. ¹H NMR (400 MHz, CDCl₃): δ 7.67–7.64 (m, aromatic-H 2H), 7.54–7.52 (m, aromatic-H 2H), 7.37 (dd, *J* = 3.2, 9.2 Hz, aromatic H, 1H), 7.11 (d, *J* = 8.4 Hz, aromatic H, 1H), 6.82 (s, 1H), 3.92, 3.89, 3.88 (s, OCH₃ × 3, 3H each); ESI MS *m/z* 313.23 (M + H)⁺.

2-(3,4-dihydroxyphenyl)-6-hydroxy-4H-chromen-4-one (9)—Yellow–brown crystalline solid. ¹H NMR (400 MHz, CDCl₃): δ 7.53 (d, *J* = 8.8 Hz, aromatic-H 1H), 7.41–7.37 (m, aromatic-H, 3H), 7.23 (dd, *J* = 2.8, 8.8 Hz, aromatic H, 1H), 6.88 (d, *J* = 8.8 Hz, aromatic H, 1H), 6.65 (s, 1H); ESI MS *m/z* 270.21 (M + H)⁺.

2-(4-hydroxy-3-methoxyphenyl)-6-methoxy-4H-chromen-4-one (10)—Yellow crystalline solid. ¹H NMR (400 MHz, CDCl₃): δ 7.62 (d, *J* = 9.2 Hz, aromatic-H 1H), 7.52 (dd, *J* = 2.4, 8.4 Hz, aromatic-H, 2H), 7.43 (d, *J* = 2.4 Hz, aromatic H, 1H), 6.88 (dd, *J* = 2.8, 8.8 Hz, aromatic-H, 1H), 7.06 (d, *J* = 8.8 Hz, aromatic-H, 1H), 6.74 (s, 1H), 3.91, 3.88 (s, OCH₃ 9 2, 3H each); ESI MS *m/z* 299.21 (M + H)⁺.

2-(3,4-dihydroxyphenyl)-6-methoxy-4H-chromen-4-one (11)—Yellow–brown crystalline solid. $^1\text{H NMR}$ (400 MHz, CDCl_3): δ 7.59 (d, $J = 9.2$ Hz, aromatic-H, 1H), 7.50 (d, $J = 2.8$ Hz, aromatic-H, 1H), 7.41 (t, $J = 2.4$ Hz, aromatic-H, 1H), 7.39 (s, aromatic-H, 1H), 7.35 (dd, $J = 3.2, 9.2$ Hz, aromatic-H, 1H), 6.88 (dd, $J = 1.2, 7.6$ Hz, aromatic H, 1H), 6.69 (s, 1H), 3.87 (s, OCH_3 , 3H); ESI MS m/z 285.19 (M + H) $^+$.

Syntheses of compounds 17–18—To a solution of naphthalen-1-ol (1 mmol) in acetone (10 mL) was added K_2CO_3 (3 eq.) followed by methyl 2-chloroacetate (1.5 eq.). The resulting mixture was heated to reflux for 20 h with TLC monitoring. The solid was filtered and the filtrate was concentrated to dryness. The residue was diluted with EtOAc and washed twice with brine. The organic portion was dried over Na_2SO_4 , filtered, and concentrated. The crude product was purified through a combiflash chromatography system with hexanes/EtOAc as eluent to afford the desired product **17**. Compound **18** was obtained by treatment of **17** (0.1 mmol) with trimethylstannanol (10 eq.) in dichloroethane (1.5 mL). The resulting mixture was heated at 80 °C for 3 h with TLC monitoring. Upon completion, the solvent was evaporated and the residue was diluted with EtOAc and washed with 5 % HCl followed by brine three times. After drying over Na_2SO_4 , the crude product was purified through a combiflash chromatography system with hexanes/EtOAc as eluent. The desired product **18** was obtained as a off-white solid.

Methyl 2-(naphthalen-1-yloxy)acetate (17)—White crystalline solid. $^1\text{H NMR}$ (400 MHz, CDCl_3): δ 8.36–8.33 (m, aromatic-H 1H), 7.80–7.77 (m, aromatic-H, 1H), 7.50–7.45 (m, aromatic H, 3H), 7.32 (t, $J = 7.6$ Hz, aromatic-H, 1H), 6.68 (d, $J = 7.6$ Hz, aromatic-H, 1H) 3.80 (s, OCH_3 , 3H); ESI MS m/z 217.21 (M + H) $^+$.

2-(naphthalen-1-yloxy)acetic acid (18)—White crystalline solid. $^1\text{H NMR}$ (400 MHz, CDCl_3): δ 8.30–8.27 (m, aromatic-H 1H), 7.77–7.75 (m, aromatic-H, 1H), 7.46–7.41 (m, aromatic H, 3H), 7.32 (t, $J = 7.6$ Hz, aromatic-H, 1H), 6.79 (d, $J = 7.6$ Hz, aromatic-H, 1H) 3.80 (s, OCH_3 , 3H); ESI MS m/z 203.19 (M + H) $^+$.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Kimberlee Moran, the Director of the Center for Forensic Science Research & Education, for her help with the manuscript preparation for the entire project. We also appreciate the help of Dr. Susan Morris-Natschke, Natural Product Research Laboratory, UNC Eshelman School of Pharmacy with proofreading and editing the manuscript. This investigation was also supported in part by NIH Grant CA 177584-01 from National Cancer Institute awarded to K.H. Lee. This study was also supported in part by the Taiwan Department of Health, China Medical University Hospital Cancer Research Center of Excellence (DOH100-TD-C-111-005).

References

1. American Cancer Society. Lifetime risk of developing or dying from cancer. Nov 14, 2013 <http://www.cancer.org/research/cancerfactsfigures/cancerfactsfigures/cancer-facts-figures-2013>. 2-1-0014

2. Maduro JH, Pras E, Willemse PH, de Vries EG. Acute and long-term toxicity following radiotherapy alone or in combination with chemotherapy for locally advanced cervical cancer. *Cancer Treat Rev.* 2003; 29:471–488. [PubMed: 14585258]
3. Kelloff GJ, Sigman CC, Greenwald P. Cancer chemoprevention: progress and promise. *Eur J Cancer.* 1999; 35:2031–2038. [PubMed: 10711244]
4. Sharma D, Sukumar S. Big punches come in nanosizes for chemoprevention. *Cancer Prev Res.* 2013; 6:1007–1010.
5. Tsao AS, Kim ES, Hong WK. Chemoprevention of cancer. *Ca-A Cancer J Clin.* 2004; 54:150–180.
6. Takemura H, Sakakibara H, Yamazaki S, Shimoi K. Breast cancer and flavonoids—a role in prevention. *Curr Pharm Des.* 2013; 19:6125–6132. [PubMed: 23448447]
7. Steward WP, Brown K. Cancer chemoprevention: a rapidly evolving field. *Br J Cancer.* 2013; 109:1–7. [PubMed: 23736035]
8. Patterson SL, Maresso KC, Hawk E. Cancer chemoprevention: successes and failures. *Clin Chem.* 2013; 59:94–101. [PubMed: 23150056]
9. Malik M, Magnuson BA. Rapid method for identification of chemopreventive compounds using multiplex RT-PCR for cyclooxygenase mRNA expression. *Cancer Detect Prev.* 2004; 28:277–282. [PubMed: 15350631]
10. Heijink DM, Fehrmann RS, de Vries EG, Koornstra JJ, Oosterhuis D, van der Zee AG, Kleibeuker JH, de Jong S. A bioinformatical and functional approach to identify novel strategies for chemoprevention of colorectal cancer. *Oncogene.* 2011; 30:2026–2036. [PubMed: 21217777]
11. Gerhauser C, Klimo K, Heiss E, Neumann I, Gamal-Eldeen A, Knauff J, Liu GY, Sitthimonchai S, Frank N. Mechanism-based in vitro screening of potential cancer chemopreventive agents. *Mutat Res.* 2003;523–524. 163–172.
12. Tokuda H, Arai T, Suzuki R, Strong JM, Schneider A, Suzuki N. Efficient evaluation of healthy tea, Gromwell seed against tumor promoting stage. *Planta Medica.* 2012; 78:1177.
13. Perestelo NR, Jimenez IA, Tokuda H, Hayashi H, Bazzocchi IL. Sesquiterpenes from *Maytenus jelskii* as potential cancer chemopreventive agents. *J Nat Prod.* 2010; 73:127–132. [PubMed: 20146433]
14. Terazawa R, Garud DR, Hamada N, Fujita Y, Itoh T, Nozawa Y, Nakane K, Deguchi T, Koketsu M, Ito M. Identification of organoselenium compounds that possess chemopreventive properties in human prostate cancer LNCaP cells. *Bioorg Med Chem.* 2010; 18:7001–7008. [PubMed: 20805033]
15. Stan SD, Singh SV. Transcriptional repression and inhibition of nuclear translocation of androgen receptor by diallyl trisulfide in human prostate cancer cells. *Clin Cancer Res.* 2009; 15:4895–4903. [PubMed: 19622577]
16. Kim Y, Kim J, Lee SM, Lee HA, Park S, Kim Y, Kim JH. Chemopreventive effects of *Rubus coreanus* Miquel on prostate cancer. *Biosci Biotechnol Biochem.* 2012; 76:737–744. [PubMed: 22484941]
17. Johnson JJ, Syed DN, Suh Y, Heren CR, Saleem M, Siddiqui IA, Mukhtar H. Disruption of androgen and estrogen receptor activity in prostate cancer by a novel dietary diterpene carnosol: implications for chemoprevention. *Cancer Prev Res (Phila).* 2010; 3:1112–1123. [PubMed: 20736335]
18. Steele VE, Sharma S, Mehta R, Elmore E, Redpath L, Rudd C, Bagheri D, Sigman CC, Kelloff GJ. Use of in vitro assays to predict the efficacy of chemopreventive agents in whole animals. *J Cell Biochem Suppl.* 1996; 26:29–53. [PubMed: 9154167]
19. Ito Y, Kawanishi M, Harayama T, Takabayashi S. Combined effect of the extracts from *Croton tiglium*, *Euphorbia lathyris* or *Euphorbia tirucalli* and n-butyrate on Epstein–Barr virus expression in human lymphoblastoid P3HR-1 and Raji cells. *Cancer Lett.* 1981; 12:175–180. [PubMed: 6266651]
20. Ito Y, Yanase S, Fujita J, Harayama T, Takashima M, Imanaka H. A short-term in vitro assay for promoter substances using human lymphoblastoid cells latently infected with Epstein–Barr virus. *Cancer Lett.* 1981; 13:29–37. [PubMed: 6272961]
21. Bertosa B, Aleksic M, Karminiski-Zamola G, Tomic S. QSAR analysis of antitumor active amides and quinolones from thiophene series. *Int J Pharm.* 2010; 394:106–114. [PubMed: 20472047]

22. Saeed BA, Saour KY, Elias RS, Al-Masoudi NA, La Cola P. Antitumor and quantitative structure activity relationship study for dihydropyridones derived from curcumin. *Am J Immun.* 2010; 6:7–10.
23. Aleksic M, Bertosa B, Nhili R, Uzelac L, Jarak I, Depauw S, vid-Cordonnier MH, Kralj M, Tomic S, Karminski-Zamola G. Novel substituted benzothiophene and thienothiophene carboxanilides and quinolones: synthesis, photochemical synthesis, DNA-binding properties, antitumor evaluation and 3D-derived QSAR analysis. *J Med Chem.* 2012; 55:5044–5060. [PubMed: 22620261]
24. Girija CR, Karunakar P, Poojari CS, Begun NS, Syed AA. Molecular docking studies of curcumin derivatives with multiple protein targets for procarcinogen activating enzyme inhibition. *J Proteom Bioinform.* 2013; 3:200–203.
25. Akihisa T, Tokuda H, Yasukawa K, Ukiya M, Kiyota A, Sakamoto N, Suzuki T, Tanabe N, Nishino H. Azaphilones, furanoisophthalides, and amino acids from the extracts of *Monascus pilosus*-fermented rice (red-mold rice) and their chemo-preventive effects. *J Agric Food Chem.* 2005; 53:562–565. [PubMed: 15686402]
26. Nakagawa-Goto K, Yamada K, Taniguchi M, Tokuda H, Lee KH. Cancer preventive agents 9. Betulinic acid derivatives as potent cancer chemopreventive agents. *Bioorg Med Chem Lett.* 2009; 19:3378–3381. [PubMed: 19481937]
27. Akihisa T, Tokuda H, Hasegawa D, Ukiya M, Kimura Y, Enjo F, Suzuki T, Nishino H. Chalcones and other compounds from the exudates of *Angelica keiskei* and their cancer chemo-preventive effects. *J Nat Prod.* 2006; 69:38–42. [PubMed: 16441065]
28. Akihisa T, Tokuda H, Ukiya M, Iizuka M, Schneider S, Ogasawara K, Mukainaka T, Iwatsuki K, Suzuki T, Nishino H. Chalcones, coumarins, and flavanones from the exudate of *Angelica keiskei* and their chemopreventive effects. *Cancer Lett.* 2003; 201:133–137. [PubMed: 14607326]
29. Sakurai N, Kozuka M, Tokuda H, Mukainaka T, Enjo F, Nishino H, Nagai M, Sakurai Y, Lee KH. Cancer preventive agents. Part 1: chemopreventive potential of cimigenol, cimigenol-3,15-dione, and related compounds. *Bioorg Med Chem.* 2005; 13:1403–1408. [PubMed: 15670948]
30. Ito C, Itoigawa M, Mishina Y, Filho VC, Enjo F, Tokuda H, Nishino H, Furukawa H. Chemical constituents of *Calophyllum brasiliense*. 2. Structure of three new coumarins and cancer chemopreventive activity of 4-substituted coumarins. *J Nat Prod.* 2003; 66:368–371. [PubMed: 12662094]
31. Suzuki M, Nakagawa-Goto K, Nakamura S, Tokuda H, Morris-Natschke SL, Kozuka M, Nishino H, Lee KH. Cancer preventive agents. Part 5. Anti-tumor-promoting effects of coumarins and related compounds on Epstein–Barr virus activation and two-stage mouse skin carcinogenesis. *Pharm Biol.* 2006; 44:178–182.
32. Akihisa T, Higo N, Tokuda H, Ukiya M, Akazawa H, Tochigi Y, Kimura Y, Suzuki T, Nishino H. Cucurbitane-type triterpenoids from the fruits of *Momordica charantia* and their cancer chemopreventive effects. *J Nat Prod.* 2007; 70:1233–1239. [PubMed: 17685651]
33. Kikuchi T, Akihisa T, Tokuda H, Ukiya M, Watanabe K, Nishino H. Cancer chemopreventive effects of cycloartane-type and related triterpenoids in in vitro and in vivo models. *J Nat Prod.* 2007; 70:918–922. [PubMed: 17503850]
34. Ito C, Itoigawa M, Mishina Y, Tomiyasu H, Litaudon M, Cosson JP, Mukainaka T, Tokuda H, Nishino H, Furukawa H. Cancer chemopreventive agents. New depsidones from *Garcinia* plants. *J Nat Prod.* 2001; 64:147–150. [PubMed: 11429990]
35. Nakagawa-Goto K, Bastow KF, Wu JH, Tokuda H, Lee KH. Total synthesis and bioactivity of unique flavone desmosdumotin B and its analogs. *Bioorg Med Chem Lett.* 2005; 15:3016–3019. [PubMed: 15913998]
36. Lin AS, Shibano M, Nakagawa-Goto K, Tokuda H, Itokawa H, Morris-Natschke SL, Lee KH. Cancer preventive agents. 7. Antitumor-promoting effects of seven active flavonolignans from milk thistle (*Silybum marianum*) on Epstein–Barr virus activation. *Pharm Biol.* 2007; 45:735–738.
37. Wang XH, Nakagawa-Goto K, Kozuka M, Tokuda H, Nishino H, Lee KH. Cancer preventive agents. Part 6: chemopreventive potential of furanocoumarins and related compounds. *Pharm Biol.* 2006; 44:116–120.
38. Cui W, Iwasa K, Tokuda H, Kashihara A, Mitani Y, Hasegawa T, Nishiyama Y, Moriyasu M, Nishino H, Hanaoka M, Mukai C, Takeda K. Potential cancer chemopreventive activity of simple

- isoquinolines, 1-benzylisoquinolines, and protoberberines. *Phytochemistry*. 2006; 67:70–79. [PubMed: 16310234]
39. Akihisa T, Takahashi A, Kikuchi T, Takagi M, Watanabe K, Fukatsu M, Fujita Y, Banno N, Tokuda H, Yasukawa K. The melanogenesis-inhibitory, anti-inflammatory, and chemo-preventive effects of limonoids in n-hexane extract of *Azadirachta indica* A. Juss. (neem) seeds. *J Oleo Sci*. 2011; 60:53–59. [PubMed: 21263200]
40. Kapadia GJ, Azuine MA, Takayasu J, Konoshima T, Takasaki M, Nishino H, Tokuda H. Inhibition of Epstein–Barr virus early antigen activation promoted by 12-O-tetradecanoylphorbol-13-acetate by the non-steroidal anti-inflammatory drugs. *Cancer Lett*. 2000; 161:221–229. [PubMed: 11090973]
41. Itoigawa M, Ito C, Tan HT, Kuchide M, Tokuda H, Nishino H, Furukawa H. Cancer chemopreventive agents, 4-phenylcoumarins from *Calophyllum inophyllum*. *Cancer Lett*. 2001; 169:15–19. [PubMed: 11410320]
42. Itoigawa M, Ito C, Tokuda H, Enjo F, Nishino H, Furukawa H. Cancer chemopreventive activity of phenylpropanoids and phytoquinoids from *Illicium* plants. *Cancer Lett*. 2004; 214:165–169. [PubMed: 15363542]
43. Ito C, Itoigawa M, Miyamoto Y, Onoda S, Rao KS, Mukainaka T, Tokuda H, Nishino H, Furukawa H. Polyprenylated benzophenones from *Garcinia assigu* and their potential cancer chemopreventive activities. *J Nat Prod*. 2003; 66:206–209. [PubMed: 12608850]
44. Ito C, Itoigawa M, Otsuka T, Tokuda H, Nishino H, Furukawa H. Constituents of *Boronia pinnata*. *J Nat Prod*. 2000; 63:1344–1348. [PubMed: 11076549]
45. Nakamura S, Kozuka M, Bastow KF, Tokuda H, Nishino H, Suzuki M, Tatsuzaki J, Morris Natschke SL, Kuo SC, Lee KH. Cancer preventive agents, part 2: synthesis and evaluation of 2-phenyl-4-quinolone and 9-oxo-9,10-dihydroacridine derivatives as novel antitumor promoters. *Bioorg Med Chem*. 2005; 13:4396–4401. [PubMed: 15914009]
46. Ito C, Itoigawa M, Kojima N, Tan HT, Takayasu J, Tokuda H, Nishino H, Furukawa H. Cancer chemopreventive activity of rotenoids from *Derris trifoliata*. *Planta Med*. 2004; 70:585–588. [PubMed: 15229812]
47. Iranshahi M, Kalategi F, Rezaee R, Shahverdi AR, Ito C, Furukawa H, Tokuda H, Itoigawa M. Cancer chemopreventive activity of terpenoid coumarins from *Ferula* species. *Planta Med*. 2008; 74:147–150. [PubMed: 18240102]
48. Akihisa T, Tabata K, Banno N, Tokuda H, Nishimura R, Nakamura Y, Kimura Y, Yasukawa K, Suzuki T. Cancer chemopreventive effects and cytotoxic activities of the triterpene acids from the resin of *Boswellia carteri*. *Biol Pharm Bull*. 2006; 29:1976–1979. [PubMed: 16946522]
49. Akihisa T, Kojima N, Kikuchi T, Yasukawa K, Tokuda H, Masters T, Manosroi A, Manosroi J. Anti-inflammatory and chemopreventive effects of triterpene cinnamates and acetates from shea fat. *J Oleo Sci*. 2010; 59:273–280. [PubMed: 20484832]
50. Takasaki M, Konoshima T, Tokuda H, Masuda K, Arai Y, Shiojima K, Ageta H. Anti-carcinogenic activity of *Taraxacum* plant. II. *Biol Pharm Bull*. 1999; 22:606–610. [PubMed: 10408235]
51. Takasaki M, Konoshima T, Tokuda H, Masuda K, Arai Y, Shiojima K, Ageta H. Anti-carcinogenic activity of *Taraxacum* plant. I. *Biol Pharm Bull*. 1999; 22:602–605. [PubMed: 10408234]
52. Sedykh A, Zhu H, Tang H, Zhang L, Richard A, Rusyn I, Tropsha A. Use of in vitro HTS-derived concentration-response data as biological descriptors improves the accuracy of QSAR models of in vivo toxicity. *Environ Health Perspect*. 2011; 119:364–370. [PubMed: 20980217]
53. Iwase Y, Takemura Y, Juichi M, Ito C, Furukawa H, Kawaii S, Yano M, Mou XY, Takayasu J, Tokuda H, Nishino H. Inhibitory effect of flavonoids from citrus plants on Epstein–Barr virus activation and two-stage carcinogenesis of skin tumors. *Cancer Lett*. 2000; 154:101–105. [PubMed: 10799745]
54. Solimeo R, Zhang J, Kim M, Sedykh A, Zhu H. Predicting chemical ocular toxicity using a combinatorial QSAR approach. *Chem Res Toxicol*. 2012; 25:2763–2769. [PubMed: 23148656]
55. Zhu H, Tropsha A, Fourches D, Varnek A, Papa E, Gramatica P, Oberg T, Dao P, Cherkasov A, Tetko IV. Combinatorial QSAR modeling of chemical toxicants tested against *Tetrahymena pyriformis*. *J Chem Inf Model*. 2008; 48:766–784. [PubMed: 18311912]

56. Zhang S, Wei L, Bastow K, Zheng W, Brossi A, Lee KH, Tropsha A. Antitumor agents 252. Application of validated QSAR models to database mining: discovery of novel tylophorine derivatives as potential anticancer agents. *J Comput Aided Mol Des.* 2007; 21:97–112. [PubMed: 17340042]
57. Medina-Franco JL, Golbraikh A, Oloff S, Castillo R, Tropsha A. Quantitative structure-activity relationship analysis of pyridinone HIV-1 reverse transcriptase inhibitors using the k nearest neighbor method and QSAR-based database mining. *J Comput Aided Mol Des.* 2005; 19:229–242. [PubMed: 16163450]
58. Shen M, Beguin C, Golbraikh A, Stables JP, Kohn H, Tropsha A. Application of predictive QSAR models to database mining: identification and experimental validation of novel anticonvulsant compounds. *J Med Chem.* 2004; 47:2356–2364. [PubMed: 15084134]
59. Zhang L, Zhu H, Oprea TI, Golbraikh A, Tropsha A. QSAR modeling of the blood-brain barrier permeability for diverse organic compounds. *Pharm Res.* 2008; 25:1902–1914. [PubMed: 18553217]
60. Zhu H, Martin TM, Ye L, Sedykh A, Young DM, Tropsha A. Quantitative structure-activity relationship modeling of rat acute toxicity by oral exposure. *Chem Res Toxicol.* 2009; 22:1913–1921. [PubMed: 19845371]
61. Greenwood, PE., Nikulin, MS. *A guide to chi squared testing.* Wiley; New York: 1996.
62. Kim M, Sedykh A, Chakravarti SK, Saiakhov RD, Zhu H. Critical evaluation of human oral bioavailability for pharmaceutical drugs by using various cheminformatics approaches. *Pharm Res.* 2014; 31:1002–1014. [PubMed: 24306326]
63. Irwin JJ, Shoichet BK. ZINC—a free database of commercially available compounds for virtual screening. *J Chem Inf Model.* 2005; 45:177–182. [PubMed: 15667143]
64. Gafner S, Lee SK, Cuendet M, Barthelemy S, Vergnes L, Labidalle S, Mehta RG, Boone CW, Pezzuto JM. Biologic evaluation of curcumin and structural derivatives in cancer chemoprevention model systems. *Phytochemistry.* 2004; 65:2849–2859. [PubMed: 15501252]
65. Shehzad A, Wahid F, Lee YS. Curcumin in cancer chemoprevention: molecular targets, pharmacokinetics, bioavailability, and clinical trials. *Arch Pharm (Weinheim).* 2010; 343:489–499. [PubMed: 20726007]
66. Meiyanto E, Hermawan A, Anindyajati A. Natural products for cancer-targeted therapy: citrus flavonoids as potent chemopreventive agents. *Asian Pac J Cancer Prev.* 2012; 13:427–436. [PubMed: 22524801]
67. Guiguemde WA, Shelat AA, Bouck D, Duffy S, Crowther GJ, Davis PH, Smithson DC, Connelly M, Clark J, Zhu F, Jimenez-Diaz MB, Martinez MS, Wilson EB, Tripathi AK, Gut J, Sharlow ER, Bathurst I, El MF, Fowble JW, Forquer I, McGinley PL, Castro S, Ngulo-Barturen I, Ferrer S, Rosenthal PJ, Derisi JL, Sullivan DJ, Lazo JS, Roos DS, Riscoe MK, Phillips MA, Rathod PK, Van Voorhis WC, Avery VM, Guy RK. Chemical genetics of *Plasmodium falciparum*. *Nature.* 2010; 465:311–315. [PubMed: 20485428]

Abbreviations

CCR	Correct classification rate
Combi-QSAR	Combinatorial quantitative structure–activity relationship
CPT	Consensus prediction thresholds
EBV-EA	Epstein–Barr virus early activation
MLR	Multiple linear regression
MOE	Molecular operating environment
PKC	Protein Kinase C

PLS	Partial least square
PTLC	Preparative thin layer chromatography
QSAR	Quantitative structure–activity relationship
RF	Random forest
SVM	Support vector machines
TLC	Thin layer chromatography
TMS	Tetramethylsilane
TPA	12- <i>O</i> -tetradecanoylphorbol-13-acetate
ZND	ZINC natural derivative

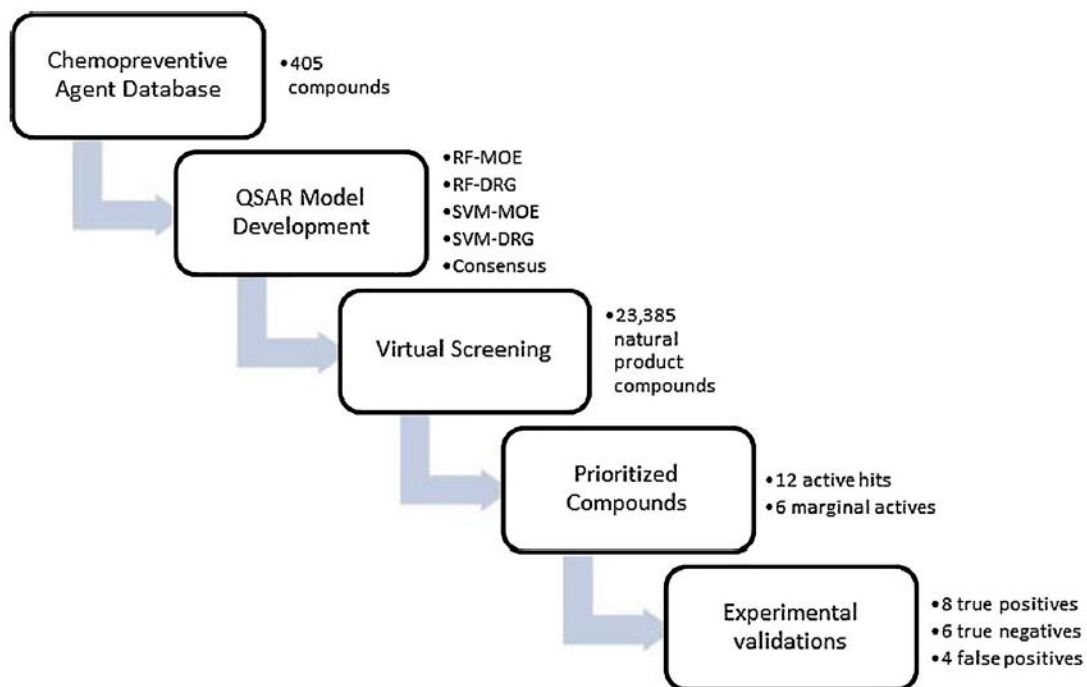


Fig. 1.
The Combi-QSAR modeling workflow of this study

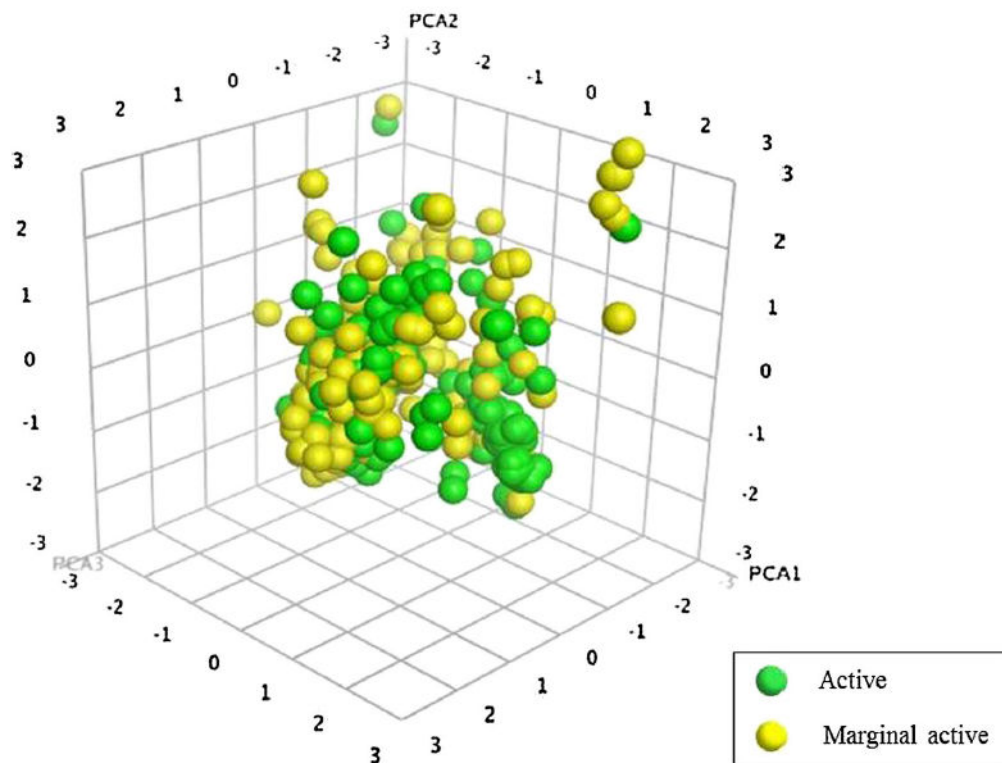


Fig. 2.
Chemical structure space of chemopreventive agent database (n = 405) using top 3 principal components of MOE descriptors

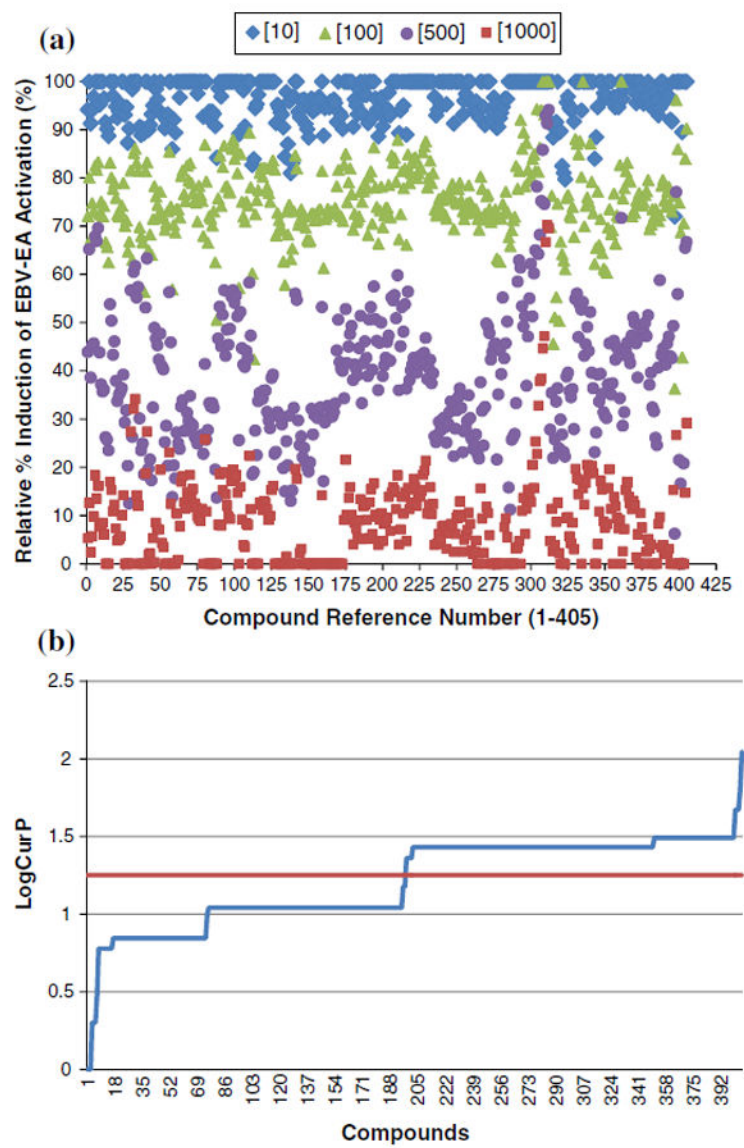


Fig. 3. The chemopreventive data obtained from the EBV-EA assays: **a** the original data shown as the distribution of relative induction of TPA-mediated EBV-EA activations at four different doses [10 (*blue*), 100 (*green*), 500 (*purple*) and 1,000 (*red*) mol ratio per 32 pmol TPA]; **b** the transformed LogCurveP results based on the four dose testing data (red line shows the active/marginally active threshold at LogCurvP = 1.25)

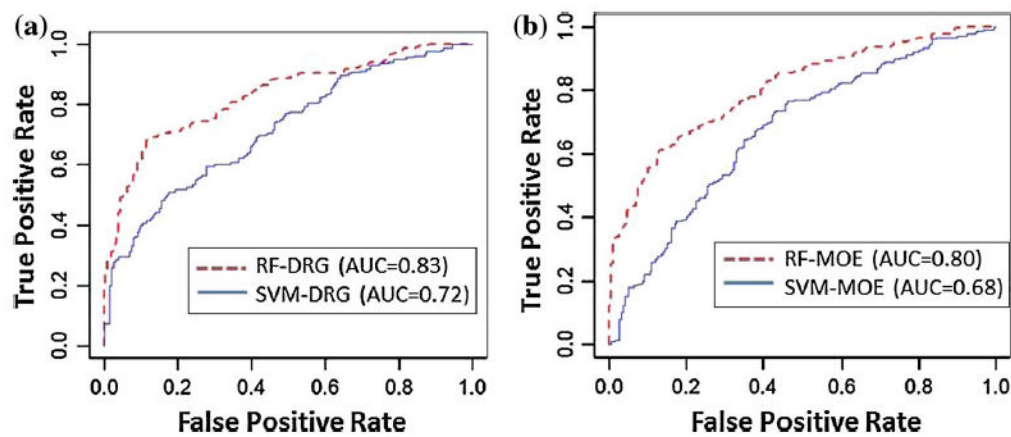


Fig. 4. ROC curves obtained as a result of five-fold cross validation: **a** two models using Dragon descriptors; **b** two models using MOE descriptors

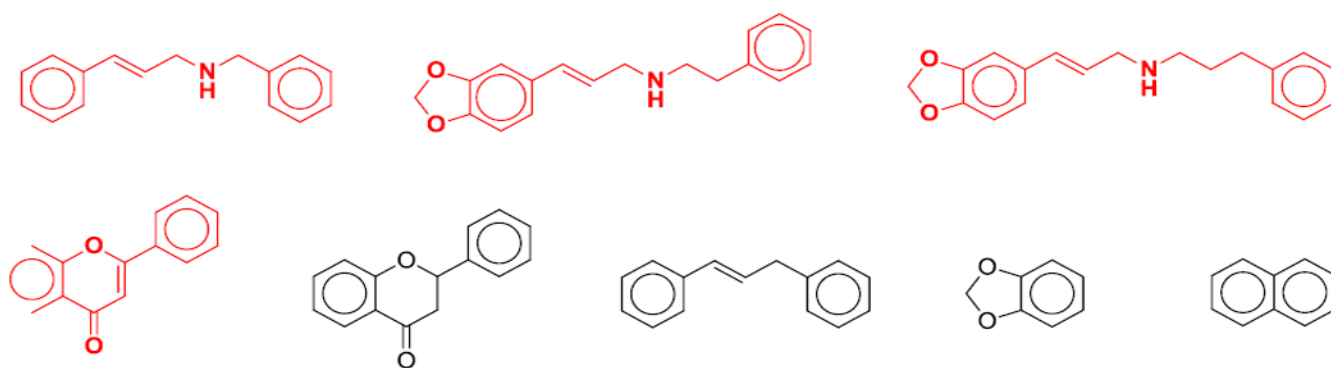
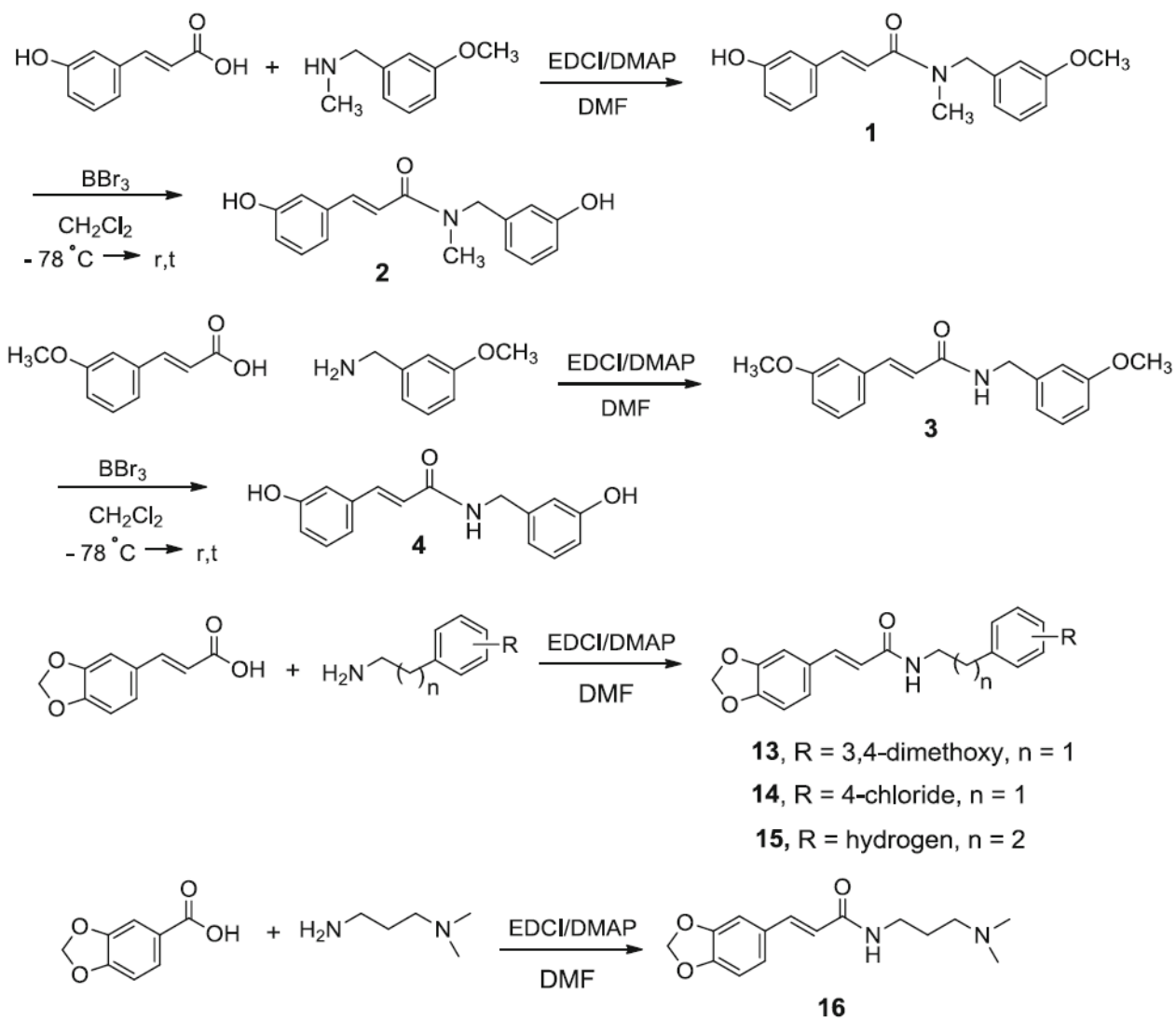
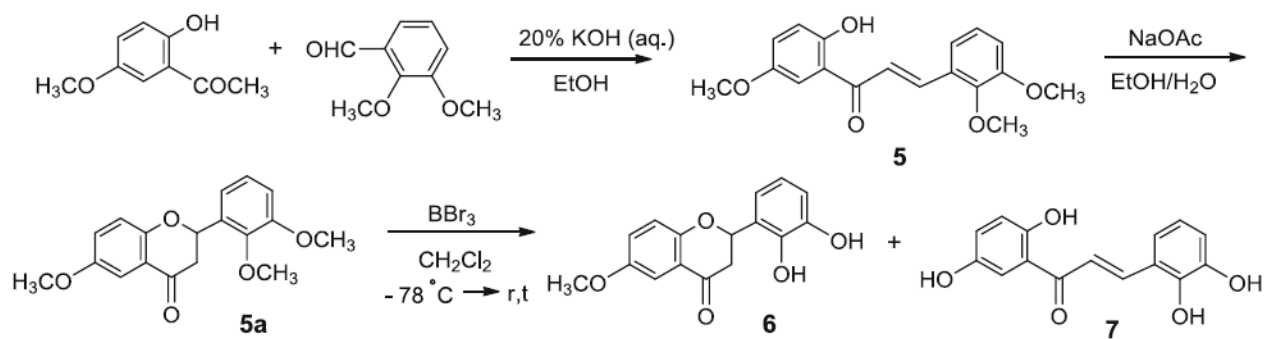


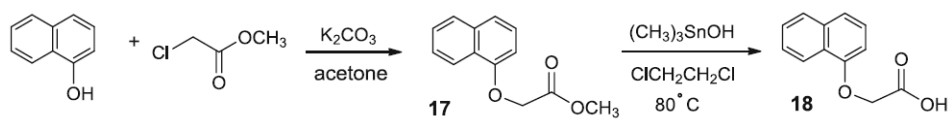
Fig. 5.
The major chemical scaffolds within the 18 new compounds. The four red ones are new scaffolds and the four black ones are within the scaffolds of the modeling set compounds



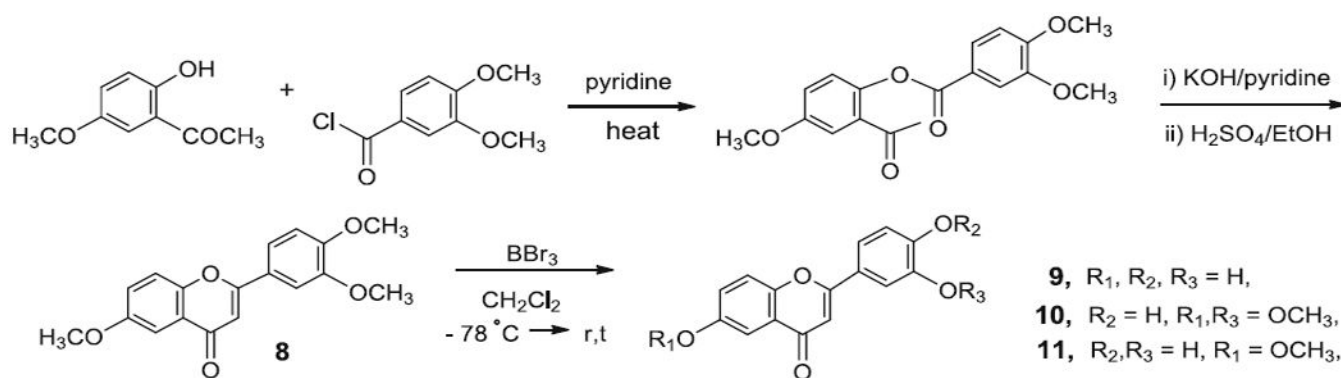
Scheme 1.
Synthesis of acrylamide substituted compounds **1–4** and **13–16**



Scheme 2.
Synthesis of flavonone **6** and its open rings compounds **5** and **7**



Scheme 3.
Synthesis of naphthalen-1-yloxy compounds **17** and **18**



Scheme 4.
Synthesis of flavones **9–11**

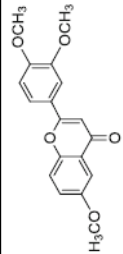
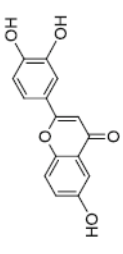
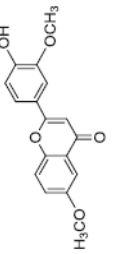
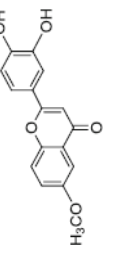
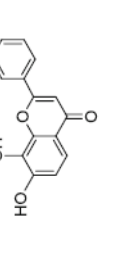
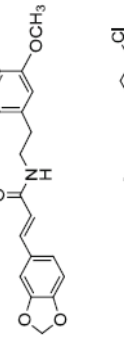
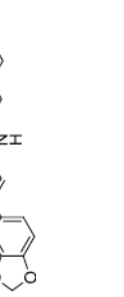
Table 1

The results of fivefold cross validation

	No CPT			CPT applied		
	Sensitivity	Specificity	CCR	Sensitivity	Specificity	CCR
RF-MOE	0.70	0.71	0.70	0.82	0.80	0.81
SVM-MOE	0.57	0.61	0.59	0.60	0.68	0.64
RF-DRG	0.75	0.74	0.74	0.82	0.80	0.81
SVM-MOE	0.66	0.63	0.65	0.80	0.74	0.77
Consensus	0.69	0.69	0.69	0.83	0.82	0.82

Table 2
 Novel chemopreventive agents identified by virtual screening and their experimental EBV-EA inhibition activities

Compounds	Responses in different concentration (nM) (folds of compound mol/TPA mol) ^a				LogCurv ^b	Pred. Act.	Exp. Act.
	32 (1,000)	16 (500)	3.2 (100)	0.32 (10)			
1	11.4 ± 0.5 (>60) ^c	47.3 ± 1.6	79.3 ± 2.5	100 ± 0.4	1.04	Active	Marginal active
2	10.3 ± 0.6 (>60)	45.3 ± 1.6	78.1 ± 2.4	100 ± 0.4	1.04	Active	Marginal active
3	13.4 ± 0.6 (>60)	48.9 ± 1.6	80.3 ± 2.5	100 ± 0.3	1.04	Active	Marginal active
4	11.6 ± 0.5 (>60)	46.1 ± 1.6	78.3 ± 2.5	100 ± 0.3	1.04	Active	Marginal active
5	6.3 ± 0.4 (>60)	42.5 ± 1.4	74.6 ± 2.3	100 ± 0.5	1.43	Active	Active
6	3.1 ± 0.4 (>60)	41.3 ± 1.3	72.8 ± 2.3	98.7 ± 0.6	1.43	Active	Active
7	4.6 ± 0.4 (>60)	40.3 ± 1.4	73.1 ± 2.4	98.6 ± 0.5	1.43	Active	Active

Compounds	Responses in different concentration (nM) (folds of compound mol/TPA mol) ^a						LogCurv ^{bb}	Pred. Act.	Exp. Act.
	32 (1,000)	16 (500)	3.2 (100)	0.32 (10)	96.5 ± 0.6	1.43			
8		3.2 ± 0.4 (>60)	32.6 ± 1.3	67.2 ± 2.3	96.5 ± 0.6	1.43	Active	Active	
9		0 ± 0.3 (>60)	26.4 ± 1.3	62.5 ± 2.3	89.1 ± 0.7	1.43	Active	Active	
10		0 ± 0.3 (>60)	29.5 ± 1.4	65.8 ± 2.4	94.6 ± 0.7	1.43	Active	Active	
11		0 ± 0.4 (>60)	28.9 ± 1.4	64.3 ± 2.5	92.4 ± 0.6	1.43	Active	Active	
12		0 ± 0.4 (>60)	29.0 ± 1.6	62.1 ± 2.5	93.3 ± 0.6	1.43	Active	Active	
13		13.1 ± 0.6 (>60)	48.6 ± 1.6	9.3 ± 2.6	100 ± 0.3	1.04	Marginal active	Marginal active	
14		17.8 ± 0.5 (>60)	51.6 ± 1.5	83.0 ± 2.4	100 ± 0.4	0.85	Marginal active	Marginal active	

Compounds	Responses in different concentration (nM) (folds of compound mol/TPA mol) ^a						LogCurv ^b	Pred. Act.	Exp. Act.
	32 (1,000)	16 (500)	3.2 (100)	0.32 (10)					
15	15.9 ± 0.5 (>60)	50.0 ± 1.5	81.3 ± 2.4	100 ± 0.3	1.04	Marginal active	Marginal active		
16	18.3 ± 0.7 (>60)	53.5 ± 1.4	84.3 ± 2.6	100 ± 0.2	0.85	Marginal active	Marginal active		
17	11.3 ± 0.5 (>60)	43.5 ± 1.5	77.3 ± 2.5	100 ± 0.4	1.04	Marginal active	Marginal active		
18	9.9 ± 0.5 (>60)	41.3 ± 1.5	75.1 ± 2.5	100 ± 0.4	1.04	Marginal active	Marginal active		
Curcumin	0 ± 0.5 (>60)	21.1 ± 1.1	80.1 ± 2.4	100 ± 0.2	1.43	-	Active		

^aTPA concentration is 32 pmol/mL

^blogCurv represents a log₁₀ transformed response fingerprint values, as described in reference 45

^cValues in parentheses represent viability percentages of Raji cells. For the determination of cytotoxicity, the cell viability is required greater than 60 %