

# Extraction and validation of substructure profiles for enriching compound libraries

Wee Kiang Yeo · Mei Lin Go · Shahul Nilar

Received: 2 May 2012 / Accepted: 7 September 2012 / Published online: 16 September 2012  
© Springer Science+Business Media B.V. 2012

**Abstract** Compounds known to be potent against a specific protein target may potentially contain a signature profile of common substructures that is highly correlated to their potency. These substructure profiles may be useful in enriching compound libraries or for prioritizing compounds against a specific protein target. With this objective in mind, a set of compounds with known potency against six selected kinases (2 each from 3 kinase families) was used to generate binary molecular fingerprints. Each fingerprint key represents a substructure that is found within a compound and the frequency with which the fingerprint occurs was then tabulated. Thereafter, a frequent pattern mining technique was applied with the aim of uncovering substructures that are not only well represented among known potent inhibitors but are also unrepresented among known inactive compounds and vice versa. Substructure profiles that are representative of potent inhibitors against each of the 3 kinase families were thus extracted. Based on our validation results, these substructure profiles demonstrated significant enrichment for highly potent compounds against their respective kinase targets. The advantages of using our approach over conventional methods in analyzing such datasets and its application in the mining of substructures for enriching compound libraries are presented.

**Keywords** Correlation rules · Substructure profiling · Kinase · Co-occurrence

## Introduction

Kinases are popular therapeutic targets, particularly for cancer-related indications [1–3]. Eukaryotic kinases are divided into eight groups based on the sequence similarity [4] of their catalytic domains, the presence of accessory domains, and their modes of regulation. These groups are TK (Tyrosine Kinase), AGC (named after the Protein Kinase A, G, and C families), CAMK (Ca<sup>2+</sup>/calmodulin-dependent kinase), CMGC (named after a set of families (CDK, MAPK, GSK3 and CLK), CK1 (Cell Kinase 1), STE (Homologs of the yeast STE7, STE11 and STE20 genes), TKL (Tyrosine Kinase-Like) and a miscellaneous group “Others”.

Achieving selectivity for kinase inhibitors remains a daunting challenge [5–14]. Given the significant amount of activity data accumulated in the published literature, it is of interest to determine if the data could be mined to yield signature profiles of substructures that are highly correlated to good potency among known kinase inhibitors. Such substructure profiles would correspond to pre-defined structural motifs that are specific to the respective kinase targets. Ligands that bind to a given kinase should share common chemical features which are absent in random compounds that do not bind to the specified kinase. The substructure profiles are anticipated to be useful in two ways: (1) They can be used to prioritize compounds that are more likely to be potent against a kinase target; (2) They can be treated as pharmacophore-like features that can be incorporated or prospectively avoided when designing novel inhibitors that interact selectively with a particular kinase target or family.

**Electronic supplementary material** The online version of this article (doi:10.1007/s10822-012-9604-8) contains supplementary material, which is available to authorized users.

W. K. Yeo (✉) · S. Nilar  
Novartis Institute for Tropical Diseases, 10 Biopolis Road,  
#05-01 Chromos, Singapore 138670, Singapore  
e-mail: weekiang@alumni.nus.edu.sg

W. K. Yeo · M. L. Go  
Department of Pharmacy, Faculty of Science, National  
University of Singapore, Block S4, 18 Science Drive 4,  
Singapore 117543, Singapore

In this paper, we apply the concept of Correlation Rules to uncover substructures that are represented with greater frequency in known potent inhibitors vis-à-vis inactive compounds and vice versa.

### Association rules, the support-confidence framework and correlation rules

Given a non-empty item-set  $I$ , an Association Rule [15] is a statement of the form  $A \rightarrow B$ , where  $A, B \subset I$  such that  $A \neq \emptyset$ ,  $B \neq \emptyset$ , and  $A \cap B = \emptyset$ . The item-set  $I$  consists of set  $A$ , called the antecedent of the Association Rule, and set  $B$  which is the consequent of the rule. Association rules provide information of co-occurrence patterns in the form of “if-then” statements, with the antecedent forming the “if” part and the consequent, the “then” part. In the context of the investigations discussed in this paper, set  $A$  consists of a binary indication of the compound’s potency while Set  $B$  indicates the presence or absence of a particular substructure motif in the compound. Association Rules are mined over a set of compounds denoted  $T$ .

$$\text{Support } s(A \rightarrow B) = P(A, B) \quad (1)$$

$$\text{Support } s(A) = P(A) \quad (2)$$

$$\text{Confidence } c(A \rightarrow B) = P(B|A) \quad (3)$$

An Association Rule has two properties that indicate its degree of uncertainty. The Support  $s(A \rightarrow B)$  (Eq. 1) of an Association Rule is the percentage of compounds in the database that contains both  $A$  and  $B$  in  $I$ .  $P(A, B)$  is the probability that  $A$  and  $B$  are simultaneously present. Hence the Support is essentially the probability that a randomly selected compound from the database will contain all items in the antecedent and the consequent, i.e. the compound will be potent as well as contain a particular substructure motif. The Confidence  $c(A \rightarrow B)$  (Eq. 3) of an Association Rule is the ratio of the Support  $s(A \rightarrow B)$  to the number of compounds that are potent (the antecedent), regardless of the presence of any particular substructure. Therefore, the Confidence is  $P(B|A)$  which is the conditional probability that a randomly selected compound will contain a particular substructure (the consequent) given that the compound is potent (the antecedent). Typically, only Rules that have Confidence and Support values above certain user-defined thresholds would be extracted. However, the Support-Confidence framework has several disadvantages and are discussed in [16, 17]. The Lift Ratio (Eq. 4) will provide a better means of assessing the strength of such rules.

For a pair of attributes [Potent, Substructure X present], the Lift Ratio is calculated as the Support value for that pair, divided by the product of the individual Support values for [Potent] and [Substructure X present] (Eq. 2).

$$\text{Lift Ratio} = \frac{s(\text{Potent} \rightarrow \text{Substructure X present})}{s(\text{Potent}) \times s(\text{Substructure X present})} \quad (4)$$

The Lift Ratio measures the extent to which the equation  $P(\text{Potent}, \text{Substructure X present}) = P(\text{Potent}) \times P(\text{Substructure X present})$  is true. If the Lift Ratio is significantly less than 1, it indicates a negative correlation between potency and the presence of Substructure X. Hence correlation is the more appropriate measure since the Support-Confidence framework has failed to detect such relationships. If the Lift Ratio equals 1, then potency and the presence of Substructure X are independent. When the Lift Ratio exceeds 1, then the larger the value, the more likely it is that potency and the presence of Substructure X in a compound is not just a chance occurrence but reflect a positive relationship between them (if compound is potent, Substructure X is also frequently present). As such, a Correlation Rule [16, 17] says that the items in an item-set are dependent. Hence the application of Correlation Rule allows us to investigate the presence of such relationships, instead of just identifying the positive relationships which are implicated by the Association Rules. Correlation Rules do not rely on the Support-Confidence framework to establish their validity. Together with the value of Lift Ratio, the statistical significance of the independence of  $A$  and  $B$  was verified by conducting the Chi-squared test for each [Potent, Substructure X present] pair.

In order to illustrate the superiority of Correlation Rules, we applied the Correlation Rules and Association Rules to the same datasets. And as described later, we found Association Rules to perform poorly. One reason for the poor outcome may be traced to the composition of the datasets where active compounds comprise only between 4.5 and 18.9 %. At these levels, the Support  $s[\text{Potent} \rightarrow \text{Substructure X present}]$  is also low. Hence using the Minimum Support of 60 % and the Confidence set at 60 %, the Association Rules approach failed to generate rules that directly associate good bioactivities with specific substructures. This is one of the shortcomings of the Association Rules that is not encountered with Correlation Rules [16, 17] which does not require the setting of pre-determined thresholds.

## Materials and methods

### Datasets

The structural and potency data found in the curated Kinase SARfari [18] dataset of the ChEMBL database [19, 20] were used. Two kinases were selected from three of the kinase groups. They are CDK2 (cyclin-dependent kinase 2) [21] and p38 alpha [22] from the CMGC group, EGFR

(epidermal growth factor receptor) [23] and SRC [24] from the TK group, AKT1 [25] and PKC $\beta$  (Protein Kinase C beta) [26] from the AGC group.

#### Data pre-processing

Only compounds with available IC<sub>50</sub> and K<sub>i</sub> values were retained. For compounds with activity data against the six selected kinases, the IC<sub>50</sub> and K<sub>i</sub> values were discretized into binary labels of 1 (active) and 0 (inactive). A compound is labeled 1 if its IC<sub>50</sub>  $\leq$  5  $\mu$ M and labeled 0 if it is  $\geq$  100  $\mu$ M. Compounds with IC<sub>50</sub> values against the six selected kinases that do not meet these threshold values are discarded. Similarly, a compound is labeled 1 if its K<sub>i</sub> value is  $\leq$  2.5  $\mu$ M and labeled 0 if it is  $\geq$  50  $\mu$ M. Compounds with IC<sub>50</sub> values between 5 and 100  $\mu$ M and K<sub>i</sub> values between 2.5 and 50  $\mu$ M are discarded to ensure that only substructures that are most likely to account for the activities will be included in the investigation.

#### Augmentation with decoy compounds

Due to the imbalance in numbers of potent compounds and inactive compounds, each of the six datasets was augmented with 8,000 identical non-redundant decoy compounds randomly selected from the ChEMBL database [19, 20]. Since these compounds are treated as putative inactives, they are labeled 0. In this report, the decoy compounds were not filtered to remove compounds that have similar physicochemical properties or structural/topological compositions. Although there are methods available to achieve these objectives [27, 28], they were not considered here because this would introduce significant bias into the decoy compounds by accentuating the difference between the decoys and the compounds being investigated in an artificial manner. This is especially so since the method used in the study is primarily based on the structural composition of the compounds. Hence the filtering process was not implemented in this investigation. Structures that

were identical to those of the active compounds were, however, removed from the list of decoy compounds and replaced with non-identical random compounds. The composition of the datasets is given in Table 1.

In order to investigate the effects of substructures, molecular fingerprints were generated using PaDEL-Descriptor [29], a software that calculates molecular descriptors and fingerprints. A total of 166 Molecular ACCess System (MACCS) fingerprint keys [30] and 881 PubChem [31, 32] fingerprint keys [33] in binary format (indicating presence or absence) were computed for each compound. In addition, for performance comparison purposes, 4,860 fingerprints as investigated by Klekota and Roth [34] were computed in binary format for each compound in the AKT1 and p38 $\alpha$  datasets.

#### Fingerprint keys selection and validation

A fivefold validation methodology was applied. Each dataset was equally and randomly divided into 6 subsets: Validation Sets 1–5 and a Test Set. At any one time only four of the five Validation Sets were used to derive the Correlation Rules. The Test Set was not used in any of the validation processes. Each of the 1,047 fingerprint keys is then used to create a contingency table for the pair-wise comparison of substructure and activity. A total of 1,047 contingency tables were generated for each kinase. The Lift Ratio was computed for each table and a Chi-square Test was conducted to establish statistical significance.

The number or frequency of compounds (F<sub>1</sub>, F<sub>2</sub>, F<sub>3</sub> and F<sub>4</sub>) that corresponds to each of the four cases is listed in Table 2. In addition, a series of Contrast scores were calculated. These scores are intended to assess the influence of the presence or absence of each Fingerprint key on the potent and non-potent compounds respectively.

$$\text{Contrast 1} = F_1 / (F_1 + F_2) \quad (5)$$

$$\text{Contrast 2} = F_2 / (F_1 + F_2) \quad (6)$$

$$\text{Contrast 3} = F_3 / (F_3 + F_4) \quad (7)$$

$$\text{Contrast 4} = F_4 / (F_3 + F_4) \quad (8)$$

$$\text{Lift Distance} = |1 - \text{Lift Ratio}| \quad (9)$$

Particular attention is paid to Contrast 1 (Eq. 5) and Contrast 4 (Eq. 8). Contrast 1 gives the proportion of actives

**Table 1** Composition of kinase datasets used in the study

Target		Number of compounds			
Group	Kinase	Active	Inactive	Augmented	Total
TK	EGFR	1,905	364	8000	10,269
	SRC	1,566	162	8000	9,728
AGC	AKT1	623	133	8000	8,756
	PKC $\beta$	396	363	8000	8,759
CMGC	CDK2	1,607	163	8000	9,770
	p38 $\alpha$	1,941	315	8000	10,256

**Table 2** An example of a contingency table for a pair-wise comparison between activity and a particular fingerprint key

	Fingerprint key present		Fingerprint key absent	
Active	F <sub>1</sub>		F <sub>2</sub>	
Inactive	F <sub>3</sub>		F <sub>4</sub>	

**Table 3** Criteria for labeling of contrast quality

	Contrast quality label			
	Excellent	Good	Moderate	Poor
Contrast 1	Both $\geq 0.8$	$0.79 < \text{Both} \leq 0.7$	$0.69 < \text{Both} \leq 0.6$	Both $< 0.6$
Contrast 4				
Contrast 2				
Contrast 3				

that have a particular fingerprint key while Contrast 4 reports on the proportion of inactives that do not have the same fingerprint key. If both Contrast 1 and Contrast 4 have values that are equal or greater than 0.8, then the potency-fingerprint key pair will have its Contrast Quality label set to “Excellent” (Table 3). If the scores are between 0.79 and 0.7, then the Contrast Quality label is set to “Good”. If their scores are between 0.69 and 0.6, it is deemed “Moderate” and if less than 0.6, it is considered “Poor”. In the same way, Contrast Quality labels (Excellent, Good, Moderate, Poor) are generated from Contrast 2 (Eq. 6) and Contrast 3 (Eq. 7). A Lift Distance score (Eq. 9, the absolute distance of the Lift Ratio from the value of 1) is then computed for each cell in the contingency table. The Lift Distance scores are summated to obtain the Total Lift Distance score, which is used to sort the Fingerprint Keys. A high Total Lift Distance score indicates that a particular Fingerprint Key is highly correlated (positively or negatively) to potency. In addition, a Chi-square test was conducted for each potency-fingerprint key pair to confirm that the pair is not independent at 95 % confidence. If the Fingerprint Key fails the Chi-square test, it is discarded.

In the final step, fingerprint keys with Contrast Qualities computed to be “Excellent”, “Good” or “Moderate” are prioritized in descending order of their Total Lift Distance scores, i.e. highest Total Lift Distance scores on top. If there are more than 10 such fingerprint keys, then only the top 10 keys with the highest Total Lift Distance are used to screen the Test Set and the five Validation Sets. If there are less than 10 such fingerprint keys, then all keys with Contrast Qualities computed to be “Excellent”, “Good” or “Moderate” are used.

### Scoring the compounds

The selected Fingerprint Keys are used to score the compounds in the respective datasets. For the screening of the Test Set and the five Validation Sets of compounds, the binary codes (indicating presence or absence) of each of the selected Fingerprint Keys are substituted by the  $(\text{Lift Distance})_{\text{present}}$  and  $-(\text{Lift Distance})_{\text{present}}$  respectively. The  $(\text{Lift Distance})_{\text{present}}$  is the summation of the Lift Distance values for all compounds containing a particular

Fingerprint key. The  $-(\text{Lift Distance})_{\text{present}}$  therefore serves as a penalty for compounds that do not contain the Fingerprint Key. Finally, for each compound in the Test Set and the remaining Validation Set, the  $(\text{Lift Distance})_{\text{present}}$  and  $-(\text{Lift Distance})_{\text{present}}$  are summed to give the Decision score. The compounds are then sorted in descending order of their Decision scores. Thereafter, the compounds are segregated into 10 deciles, each containing a similar number of compounds. Decile 1 will have compounds with the highest Decision score (high probability that these compounds will be potent) and Decile 10 comprise compounds with the lowest Decision score (low probability that these compounds will be potent).

### Performance evaluation

Predictive accuracy measures are often used to evaluate the performance of classifiers [35]. However, for methods based on scoring, this is inadequate for the following reasons: First, predictive accuracy measures typically compute the percentage of compounds that are classified correctly (or erroneously) by the classifier. This is inappropriate for the evaluation of the Decision score since the scoring procedure does not have the concept of correct or incorrect classification. Each compound is merely assigned a numerical value that indicates the likelihood of it belonging to the active class. Second, predictive accuracy generally does not provide adequate indication of the distribution of the active compounds in the top deciles and therefore are unable to evaluate the Decision score effectively. In view of these limitations, an Enrichment Factor was introduced to evaluate performance. The Enrichment Factor gives the cumulative percentage of potent compounds from Decile 1 to the next decile compared to the random distribution of the potent compounds under similar conditions when the Correlation Rules are not applied. If the cumulative percentage of active compounds in the top three deciles exceeds 70 %, the Decision score is deemed to be capable of selecting and prioritizing active compounds embedded within a large dataset.

The performance of the Correlation Rules approach was also compared to the classical similarity search using the same 1,047 fingerprint keys. The size and graphical atomic extent of the fingerprints used in this study can be viewed

as having less chemical relevance and are more promiscuous in picking up compounds of interest. This concern was addressed by analyzing two of the data sets using a different set of 4,860 fingerprint keys [34] containing more chemically-meaningful fragment sizes as a comparison.

#### Diversity of the datasets

The similarity between the active compounds in each dataset and the compounds in the reference set (inactive and augmented sets of compounds) was investigated by computing the mean Tanimoto coefficient [36] scores. These scores were obtained by first calculating the Tanimoto coefficient of each active compound against each reference set compound in a pair-wise manner based on the same 1,047 fingerprint keys used for elucidating the substructure profiles. Thereafter, the averages of these Tanimoto coefficient scores were calculated. These scores were then visualized using box-and-whisker plots (Fig. 2).

## Results and discussion

#### Fingerprint keys selection

For every dataset, each of the Fingerprint keys was scored in a pair-wise manner (Potency-Fingerprint Key) in order to ascertain its level of correlation with potency. Due to space constraints, the full set of results for the 1,047 pair-wise comparisons from all 6 datasets are not included in this manuscript but are available from the authors on request. Tables 4, 5, 6, 7, 8, 9 record the graphical depictions of the 10 top-scoring fingerprint keys from the corresponding Validation Set 1 and Test Set of each dataset. Attention should be drawn to PKC $\beta$  Validation Sets 4 and 5 (Table S1, Supplementary Information) as they contain only 9 selected Fingerprint Keys. This is because only 9 Fingerprint Keys received the Moderate label or better in terms of Fingerprint Quality.

An inspection of the top three scoring fingerprint keys shows a recurring preference for nitrogen (N)-containing entities among the TK and CMGC families. Notably, the 2-nitrogen fingerprint PubchemFP621 is found in the 1st Decile of EGFR (TK), SRC (TK) and CDK2 (CMGC). Another 2-nitrogen fingerprint PubchemFP674 occurs in the 1st Decile of p38 $\alpha$  (CMGC). The 2nd and 3rd Deciles of the TK and CMGC families are similarly populated by nitrogen containing fingerprints. In contrast, nitrogen containing fingerprints are not explicitly recognized in the top three deciles of the AGC kinases. For this family, the top scoring fingerprints are those that embed ring structures. It is difficult to interpret the significance of these observations as the fingerprints are not discrete entities, namely that one fingerprint may overlap with another.

#### Validation

The prediction results of the six datasets were evaluated using fivefold cross validation and external validation. The results are given in Fig. 1.

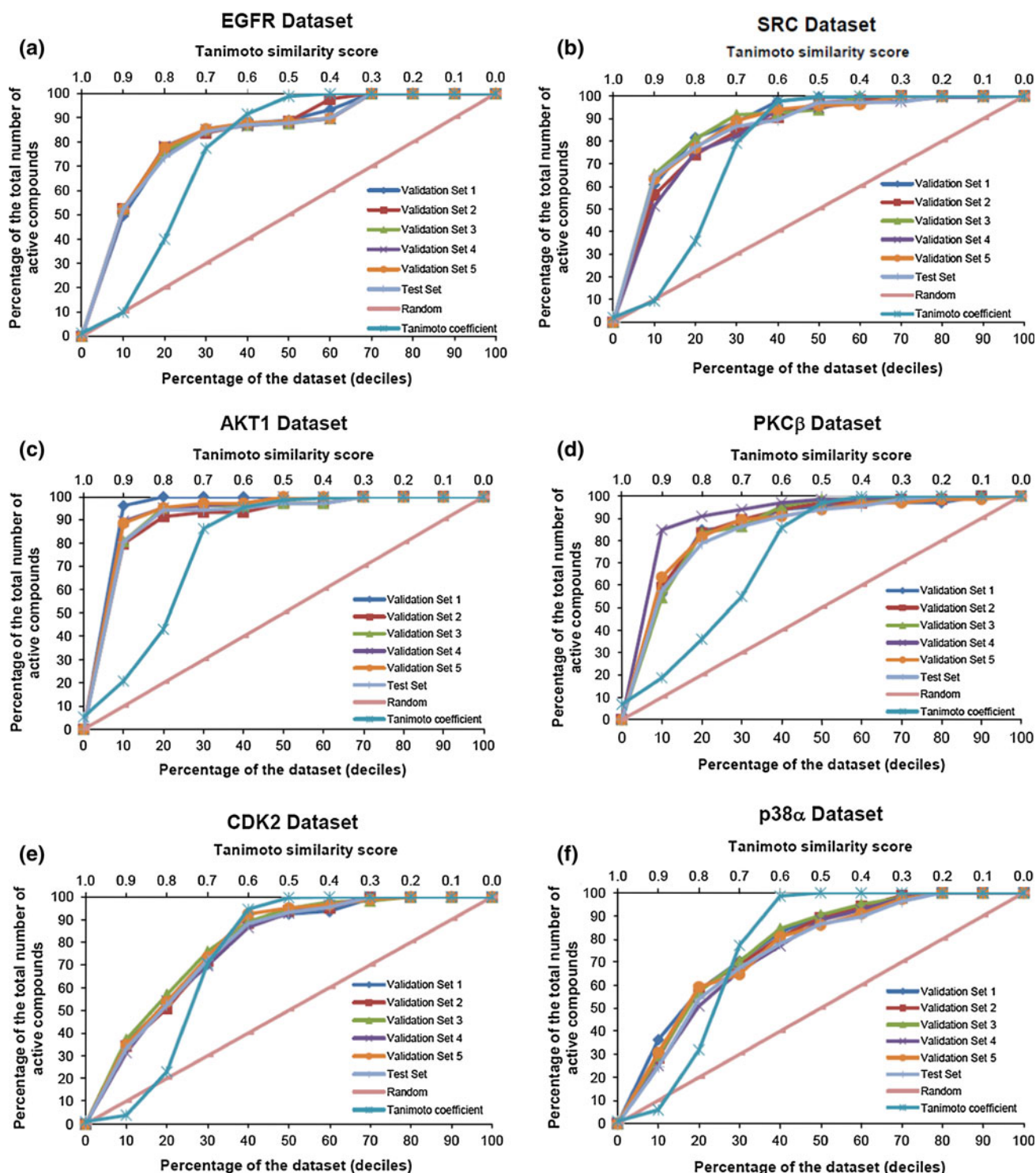
For the EGFR dataset, slightly more than half of the potent compounds in Validation Sets 1–5 (average of 51.6 %) were found within the 1st Decile (Fig. 1a). Compared to the random selection of actives using the Correlation Rules, this method outperforms by approximately 5 times (Fig. 1a). By the 3rd Decile (i.e. top 30 % of each Validation Set), an average of 84.9 % of the potent compounds have been picked up and by the 7th Decile, all the potent compounds in each Validation Set have been successfully identified. The external validation of the EGFR Test Set gave comparable results. Equally good outcomes were obtained with the SRC dataset (Fig. 1b) where nearly 87 % of the potent compounds in the Validation Sets were identified by the 3rd Decile. The external validation of the SRC Test Set produced comparable results. In the same way, 63.6 % of actives (average of 5 Validation Sets) in the PKC $\beta$  dataset (Fig. 1d) were identified in the 1st Decile and 89.1 % in the 3rd Decile. The external validation of the PKC $\beta$  Test Set gave comparable results. Of the 5 PKC $\beta$  Validation Sets, the best outcome was obtained with Validation Set 4 where 84.8 % of actives were picked up in the 1st Decile.

Even more encouraging results were obtained with the AKT1 dataset (Fig. 1c). Here, an average of 87 % of actives were captured in the 1st Decile and by the 3rd Decile, 96.5 % of the actives were identified. On inspection, it was found that the AKT1 Validation Set 1 outperformed the other Validation/Test Sets, with 96.1 % of actives captured in the 1st Decile and all actives (100 %) by the 2nd Decile.

In contrast to the other kinase datasets, the actives in the CDK2 and p38 $\alpha$  datasets were less readily picked up by the Correlation Rules. Only 34.8 % of actives were identified in the 1st Decile for the CDK2 Validation Sets although the situation improved somewhat by the 3rd Decile with 78.3 % actives successfully identified. In the case of the p38 $\alpha$  dataset, an average of 56.6 % of the actives was identified by the 3rd Decile which compared poorly with the other kinase datasets. Interestingly, both CDK2 and p38 $\alpha$  are CMGC kinases and the low levels to which their datasets were enriched with actives suggest that the method may discriminate between different kinase families. Follow-up investigations on other members of the different kinase families would be helpful in this regard.

In Fig. 1, for each dataset, one corresponding Tanimoto coefficient curve was plotted on a secondary horizontal axis. However, it must be emphasized that the Tanimoto coefficient curve in each panel should not be directly





**Fig. 1** Enrichment curves of the fivefold cross validation results using the respective datasets. The plots show the cumulative percentage of the active compounds at each decile (the primary horizontal axes) when the correlation rules approach was used. As reference, the plots on the secondary horizontal axes show the

compared to either the enrichment curves or the random diagonal curve. The enrichment curves depict the cumulative percentage of active compounds in each dataset that

cumulative percentage of the active compounds at each bin of the Tanimoto coefficient scores when the classical similarity search approach was used. However, direct comparison was not possible between the enrichment curves and the Tanimoto coefficient

was selected by each decile (primary horizontal axis) in descending order of priority using the Correlation Rules approach. On the other hand, the Tanimoto curve of each

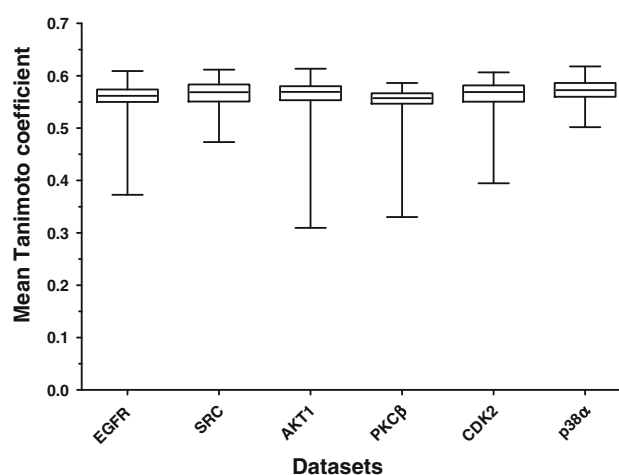
dataset plots the cumulative percentage of active compounds that was selected at a particular Tanimoto coefficient threshold between each active compound and all other active compounds in that dataset. Based on Fig. 1, there were no obvious trends or correlations between the percentages of active compounds selected based on the Tanimoto coefficient thresholds and the kinase family from which the datasets were derived or the level of enrichment achieved by the Correlation Rules approach. A detailed discussion of the observations arising from the Tanimoto curves can be found under the “Performance comparison” sub-section. The same Tanimoto curves were also depicted in Fig. 3 in order to facilitate comparison across the six datasets.

Notwithstanding the encouraging results, there are limitations in our approach. One limitation is that the comparisons were conducted in a pair-wise manner, i.e. each Fingerprint Key was independently compared to potency. This approach may increase the likelihood of selecting collinear Fingerprint Keys as part of the substructure profile. However, the alternative approach is to take into account multiple Fingerprint Keys at the same time but this will lead to a network of significantly more complex many-to-many relationships between the 1,047 Fingerprint Keys. Hence the current approach remains a viable option.

The other limitation is that the substructures defined by the Fingerprint Keys may be too small in size to be useful to medicinal chemists for lead optimization. However, if the selected substructures are treated as a signature profile for a specific protein target, it can still be useful for enriching compound libraries or for prioritizing compounds. Most of the top 10 selected Fingerprint Keys remained stable within each dataset. There was also no significant overlap between the selected Fingerprint Keys across datasets. This indicates that our approach is able to pick out unique substructure profiles even among members of the same kinase group. The ability to distinguish between compounds designed to act on kinases from the same family may serve the useful purpose of highlighting compounds that are more selective. Alternative fragmentation methods may also be employed to define the substructures. In particular, methods that give rise to entities that fall within the size range of fragments may be interesting to medicinal chemists, even though Klekota-Roth fingerprint keys have been explored with unsatisfactory results as described below. Another approach is to apply the method to extract substructures from fragment-based screening data instead of data from complete compounds as we have done here. The dependence of the method on the number and structural diversity of the augmentation compounds is another area that needs further consideration. Clearly, these are areas that could be explored in greater detail.

### Diversity of the datasets

The box-and-whisker plots for all the six datasets are shown in Fig. 2. It is observed that the inter-quartile ranges of the mean Tanimoto coefficient scores of the datasets fall between 0.54 and 0.59. The narrow inter-quartile range indicates that the active compounds are structurally more similar among themselves as compared to compounds in the reference set (inactive and augmented compounds). Moreover, since the comparison in the box-and-whisker plots is made against the reference set, the corollary is that the active and reference compounds are to a certain extent, structurally dissimilar. The structural similarity among the active compounds may reflect the fact that compounds intended as kinase inhibitors tend to be rationally designed and thus, likely to contain similar motifs or scaffolds. In addition, the curated data found in the ChEMBL Kinase SARfari database contain mostly compounds from analogue series extracted from medicinal chemistry literature rather than compounds with diverse scaffolds. In contrast, the 8,000 augmentation compounds were randomly selected without prior filtering by physico-chemical or structural/topological criteria. Consequently, their mean Tanimoto coefficient scores would be relatively higher, as observed in Fig. 2. A more elaborate filtering scheme such as those adopted by others [27, 28] may give rise to lower mean Tanimoto coefficient scores but this would be achieved at the expense of artificially enhancing structural diversity to some degree. We have shown here that even without artificially enhancing structural diversity, promising enrichment results could be achieved by applying



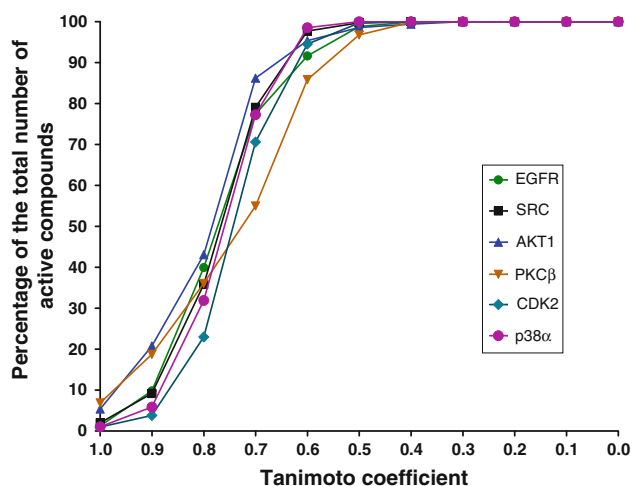
**Fig. 2** Box-and-whisker plots of the mean Tanimoto coefficient scores of the active compounds in each dataset when compared against the inactive compounds and augmented compounds. Ends of the *whiskers* represent the minimum and maximum mean Tanimoto coefficient scores of all the compounds in each dataset

Correlation Rules for the extraction of substructure profiles.

It can be seen from Table 1 that ratios of active compounds to total number of inactive and augmented compounds vary among the six datasets. Notably, ratios for the AGC kinases AKT1 and PKC $\beta$  are large, at 1:13 and 1:21 respectively as compared to the ratios for the CMGC kinases (CDK2 1:5; p38 $\alpha$  1:4) and TK kinases (EGFR 1:4; SRC 1:5). Coincidentally, the AGC kinases showed the best enrichment profiles (Fig. 1) which suggests that increasing the number of augmentation compounds to the extent found in the AGC kinase datasets may lead to better enrichment performance.

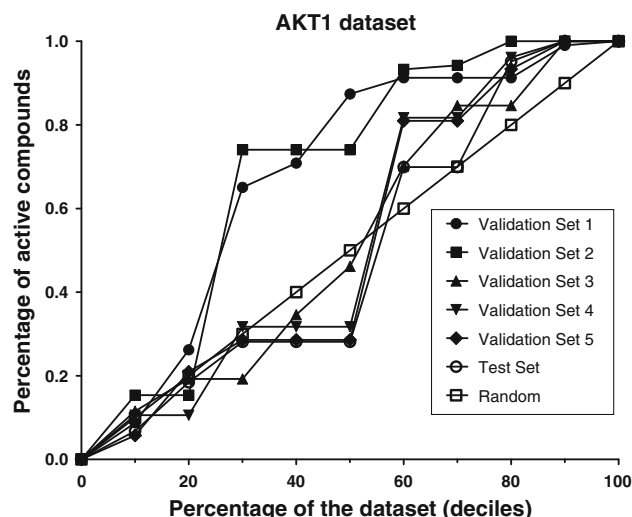
### Performance comparison

The consequences of using a classical similarity search method were investigated as a comparison. The questions we have tried to answer are: (1) if a known active compound is used as a similarity matching template to select compounds from a retrospective database, how many of those compounds picked up based on high structural similarity would be a known active compound and (2) the minimum level of structural similarity required in order to pick up such active compounds. This involved computing the Tanimoto coefficient scores between each active compound and all the other active compounds in each dataset. The Tanimoto coefficient scores were generated based on similarity matching using the same 166 MACCS and 881 PubChem fingerprint keys. In this case, however, no selection or filtering was done to the fingerprints, i.e. all 1,047 fingerprint keys were used for similarity matching.

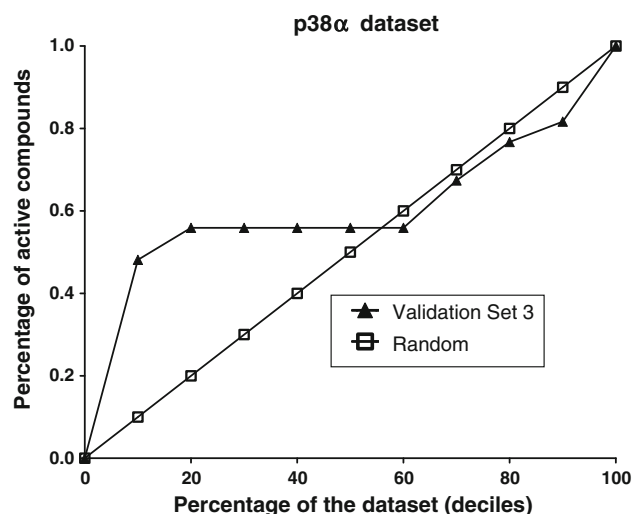


**Fig. 3** The plots show the cumulative percentage of the active compounds at each bin of the Tanimoto coefficient scores. The Tanimoto coefficient scores were calculated based on the comparison of each active compound against all other active compounds in each dataset

In conducting a similarity search, it is necessary to use a cut-off value above which, molecules are considered to be similar. This value is subjective and the retrieval of the number of similar compounds (or active compounds in this application) is dependent on this value. The results (Fig. 3) show that using any one of the active compounds as the template for similarity matching at Tanimoto coefficient score  $>0.7$  is unlikely to pick up more than 44 % of the active compounds in each dataset. Moreover, the classical similarity search method is unable to assign a priority or importance to each compound that matches the template



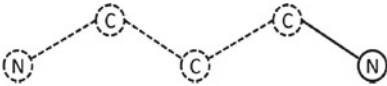
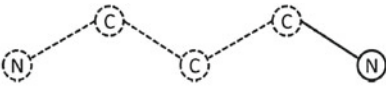
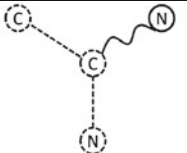
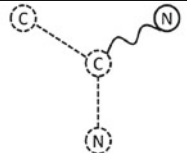
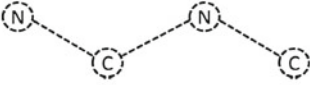

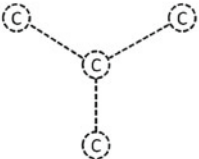
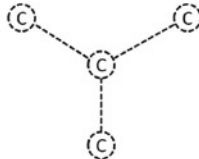
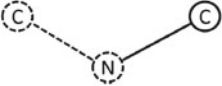
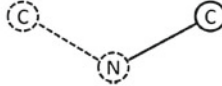
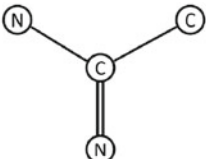
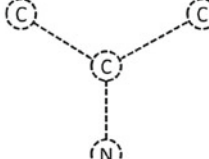
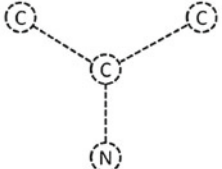
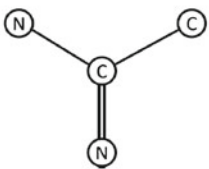
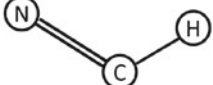
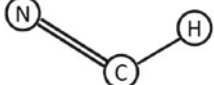
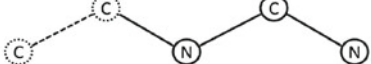
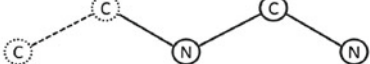
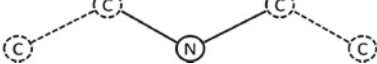
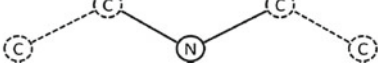
**Fig. 4** Enrichment curve of the fivefold cross validation results using the AKT1 dataset derived from the Klekota-Roth fingerprint keys. The plots show the cumulative percentage of the active compounds at each decile



**Fig. 5** Enrichment curve of the validation results using the p38 $\alpha$  dataset validation Set 3 derived from the Klekota-Roth fingerprint keys. The plots show the cumulative percentage of the active compounds at each decile

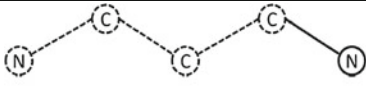
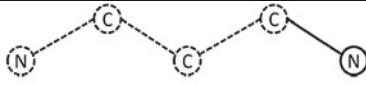
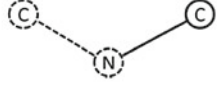
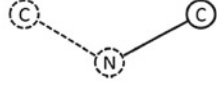
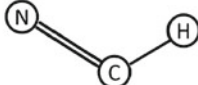
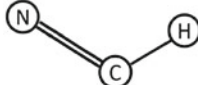
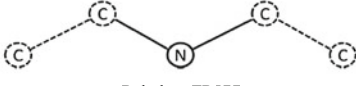
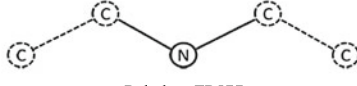
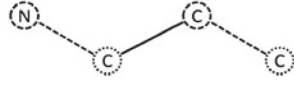
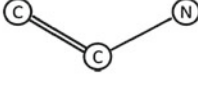
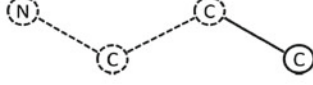

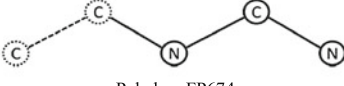
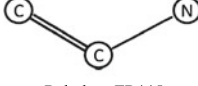
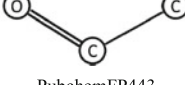


**Table 4** The top 10 fingerprint keys of the EGFR validation and test sets selected by the scoring scheme

	EGFR Validation Set 1	EGFR Test Set
1	 PubchemFP621	 PubchemFP621
2	 PubchemFP378	 PubchemFP378
3	 PubchemFP572	 PubchemFP572
4	 PubchemFP385	 PubchemFP385
5	 PubchemFP491	 PubchemFP491
6	 PubchemFP438	 PubchemFP386
7	 PubchemFP386	 PubchemFP438
8	 PubchemFP447	 PubchemFP447
9	 PubchemFP674	 PubchemFP674
10	 PubchemFP577	 PubchemFP577

Dashed lines and circles denote aromatic bonds and atoms respectively. Continuous lines and circles denote aliphatic bonds and atoms respectively. Curly lines denote any bond type and the question mark denote any atom

**Table 5** The top 10 fingerprint keys of the SRC validation and test sets selected by the scoring scheme

	SRC Validation Set 1	SRC Test Set
1	 PubchemFP621	 PubchemFP621
2	 PubchemFP491	 PubchemFP491
3	 PubchemFP447	 PubchemFP447
4	 PubchemFP577	 PubchemFP577
5	atom with a charge of not +0 MACCSFP49	atom with a charge of not +0 MACCSFP49
6	$\geq 4$ N PubchemFP16	 PubchemFP530
7	 PubchemFP445	 PubchemFP523
8	 PubchemFP484	$\geq 4$ N PubchemFP16
9	 PubchemFP674	 PubchemFP445
10	 PubchemFP443	$\geq 1$ saturated or aromatic nitrogen-containing ring size 6 PubchemFP180

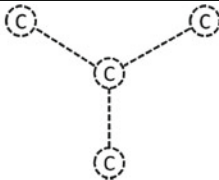


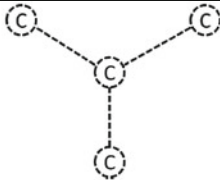
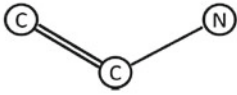
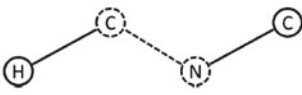
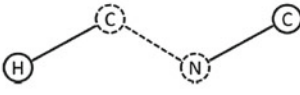
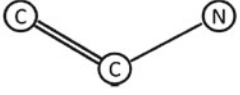
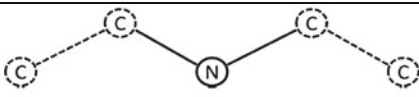
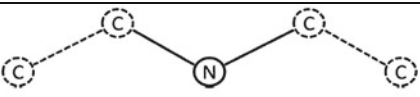

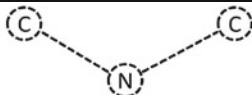
Dashed lines and circles denote aromatic bonds and atoms respectively. Continuous lines and circles denote aliphatic bonds and atoms respectively. Curly lines denote any bond type and the question mark denote any atom

compound at a specific Tanimoto coefficient score. Testing compounds thus mined, above a certain similarity score may not necessarily lead to better actives due to the lack of priority assignment. The other disadvantage of the classical similarity search method is that it is only capable of picking up compounds that are structural analogues of the template compound (i.e. highly similar scaffolds) since all 1047 fingerprint keys were used. This may not be useful for finding novel scaffolds in the databases made up of diverse scaffolds.

The size and graphical atomic extent of the 1,047 fingerprint keys used in this study can be viewed as having less chemical relevance and greater promiscuity in picking

up compounds of interest. To address this concern, we analyzed two of the data sets using a different set of 4,860 fingerprint keys reported by Klekota and Roth [34] which contain more chemically-meaningful fragment sizes. Using the Klekota-Roth fingerprint keys in the Correlation Rules approach, it was observed (Fig. 4) that the enrichment performance for the AKT1 dataset was adversely affected. One reason may be that the Klekota-Roth fingerprint keys with acceptable Contrast Quality were significantly less than the ones from MACCS-PubChem. For AKT1 Validation Sets 4, 5 and Test Set, there were only 4 Klekota-Roth fingerprint keys with acceptable Contrast Quality. Validation Set 2 had 5 such fingerprint keys while

**Table 6** The top 10 fingerprint keys of the AKT1 validation and test sets selected by the scoring scheme

	AKT1 Validation Set 1	AKT1 Test Set
1	atom with a charge of not +0 MACCSFP49	atom with a charge of not +0 MACCSFP49
2	$\geq 4$ aromatic rings PubchemFP261	$\geq 4$ aromatic rings PubchemFP261
3	$\geq 2$ hetero-aromatic rings PubchemFP258	$\geq 2$ hetero-aromatic rings PubchemFP258
4	 PubchemFP385	 PubchemFP372
5	 PubchemFP372	 PubchemFP385
6	 PubchemFP445	 PubchemFP491
7	 PubchemFP491	 PubchemFP445
8	 PubchemFP577	 PubchemFP577
9	$\geq 4 N$ PubchemFP16	$\geq 4 N$ PubchemFP16
10	 MACCSFP154	 PubchemFP403



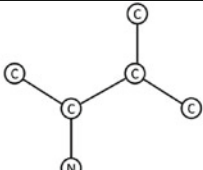
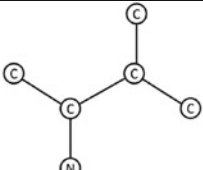


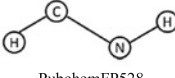
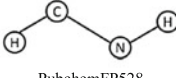
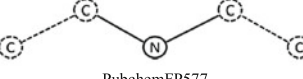
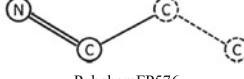
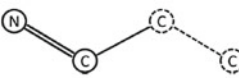
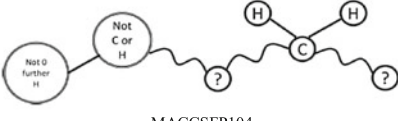
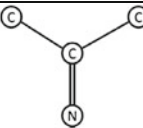
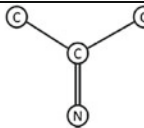
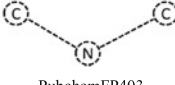

Dashed lines and circles denote aromatic bonds and atoms respectively. Continuous lines and circles denote aliphatic bonds and atoms respectively. Curly lines denote any bond type and the question mark denote any atom

Validation Sets 1 and 3 had 8 fingerprint keys. This is despite the fact that there were four times as many fingerprint keys to choose from compared to the ones from MACCS-PubChem. The variable number of the Klekota-

Roth fingerprint keys across the Validation and Test Sets may also indicate instability.

The results from the p38x dataset were more appalling. All Validation Sets, with the exception of Validation Set 3,

**Table 7** The top 10 fingerprint keys of the PKC $\beta$  validation and test sets selected by the scoring scheme

	PKC $\beta$ Validation Set 1	PKC $\beta$ Test Set
1	 MACCSFP89	 MACCSFP89
2	$\geq 2$ any ring size 5 PubchemFP150	$\geq 2$ any ring size 5 PubchemFP150
3	 PubchemFP712	 PubchemFP712
4	 PubchemFP597	 PubchemFP597
5	 PubchemFP528	 PubchemFP528
6	 PubchemFP577	 PubchemFP576
7	 PubchemFP576	$\geq 1$ saturated or aromatic nitrogen-containing ring size 5 PubchemFP145
8	$\geq 1$ saturated or aromatic nitrogen-containing ring size 5 PubchemFP145	 MACCSFP104
9	 PubchemFP431	 PubchemFP431
10	 PubchemFP403	 PubchemFP403

Dashed lines and circles denote aromatic bonds and atoms respectively. Continuous lines and circles denote aliphatic bonds and atoms respectively. Curly lines denote any bond type and the question mark denote any atom

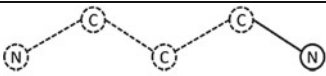

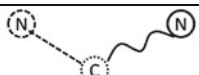

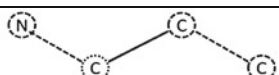
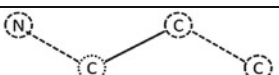
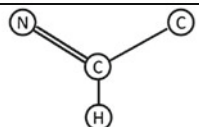
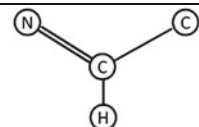
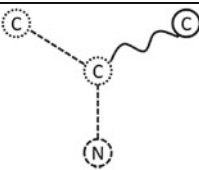
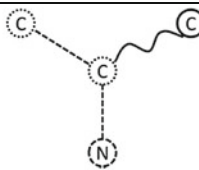
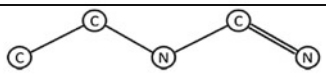
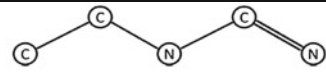
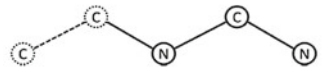
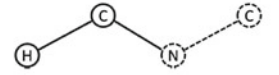
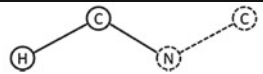
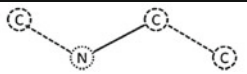
selected only 1 Klekota-Roth fingerprint key with sufficient Contrast Quality. Not surprisingly, this had a detrimental impact on the enrichment performance. Since those Validation Sets and Test Set used only 1 Klekota-Roth fingerprint key, the assignment of priority was dependent on the presence or absence of fingerprint key KR3956 (the C–O aliphatic bond), which made it impossible to plot an enrichment curve. Therefore, there are only two orders of priority for such cases. On average, only 33 % of the active compounds in these Validation Sets and Test Set have been given top priority status compared to an average of 54 % in

the top decile using the MACCS-PubChem fingerprint keys. For the p38 $\alpha$  Validation Set 3 (Fig. 5), only 55.9 % of the active compounds were selected even after 60 % of the p38 $\alpha$  dataset has been selected.

The results showed that applying the Correlation Rules to the Klekota-Roth privileged substructures did not enrich the selected kinase libraries to the same extent as the 166 MACCS and 881 PubChem fingerprints. Larger fragments, unlike smaller fragments, are possibly less capable of discerning the subtle differences between active and inactive compounds. The other observation was that it was



**Table 8** The top 10 fingerprint keys of the CDK2 validation and test sets selected by the scoring scheme

	CDK2 Validation Set 1	CDK2 Test Set
1	 PubchemFP621	 PubchemFP621
2	 PubchemFP379	 PubchemFP379
3	 PubchemFP530	 PubchemFP530
4	$\geq 1$ saturated or aromatic nitrogen-containing ring size 5 PubchemFP145	$\geq 1$ saturated or aromatic nitrogen-containing ring size 5 PubchemFP145
5	 PubchemFP435	 PubchemFP435
6	 PubchemFP357	 PubchemFP357
7	 PubchemFP596	$\geq 4 N$ PubchemFP16
8	$\geq 4 N$ PubchemFP16	 PubchemFP596
9	 PubchemFP674	 PubchemFP521
10	 PubchemFP521	 PubchemFP482

Dashed lines and circles denote aromatic bonds and atoms respectively. Continuous lines and circles denote aliphatic bonds and atoms respectively. Curly lines denote any bond type and the question mark denote any atom

harder to match the larger fragments thus resulting in a large majority of the Klekota-Roth fingerprint keys being absent in the datasets investigated. The small fragments used in the MACCS-PubChem substructure profiles as illustrated in Tables 4, 5, 6, 7, 8, 9 were intended to provide a complete profile, not as individual substructures to be employed as cues for lead optimization.

Several practical applications may be proposed for the Correlation Rules approach described here. An obvious application is the follow-up of high-throughput screening (HTS) which currently involves substructure and similarity searching using HTS hits or other known actives as

template compounds. It may be applied to mine the noise in a HTS screen for false negatives or as an additional prioritization step in a typical virtual screening workflow. The substructure profiles can be used to virtually screen public or corporate compound databases during the hit identification phase. Selected compounds will subsequently be docked against the therapeutic target of interest. The advantages of the Correlation Rules are that the shortlisted compounds are ranked even before they have been assayed or docked, and that the ranking is not based on gross structural similarity as would be the case for a classical similarity approach, but on fingerprints that populate

**Table 9** The top 10 fingerprint keys of the p38 $\alpha$  validation and test sets selected by the scoring scheme

	p38 $\alpha$ Validation Set 1	p38 $\alpha$ Test Set
1	 PubchemFP674	 PubchemFP674
2	 PubchemFP373	 PubchemFP373
3	 PubchemFP372	 PubchemFP372
4	 PubchemFP435	 PubchemFP435
5	 PubchemFP445	 PubchemFP445
6	 PubchemFP379	 PubchemFP379
7	 PubchemFP491	 PubchemFP491
8	 MACCSFP62	 MACCSFP62
9	 PubchemFP636	 PubchemFP577
10	 MACCSFP107	 PubchemFP636

Dashed lines and circles denote aromatic bonds and atoms respectively. Continuous lines and circles denote aliphatic bonds and atoms respectively. Curly lines denote any bond type and the question mark denote any atom

known active compounds for a said therapeutic target. The approach may also be applied to the derivation of substructure profiles that are representative of compounds known to have specific toxicity liabilities. This may help to flag potentially worrisome compounds in virtual compound libraries even before they have been synthesized.

## Conclusion

In this paper we have proposed a novel application of Correlation Rules to extract substructure profiles that are highly correlated to the potency of compounds known to be active against a specific protein target. Comparisons with Association Rules and classical similarity search have been performed and the superiority of the Correlation Rules has been established for six kinase datasets. Furthermore, a more chemically-meaningful fingerprint set was used to compare the success of the method and the dependence of the results on the fingerprints used. Taken together, the substructure profiles extracted in this work may be used to enrich compound libraries or to prioritize compounds for synthesis or evaluation against a specific therapeutic target.

**Acknowledgments** The PhD scholarship awarded to WKY from the Novartis Institute for Tropical Diseases is gratefully acknowledged.

## References

- Zhang C, Habets G, Bollag G (2011) *Nat Biotechnol* 29(11):981
- Eglen R, Reisine T (2011) *Pharmacol Ther* 130(2):144
- Eglen RM, Reisine T (2009) *Assay Drug Dev Technol* 7(1):22
- Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S (2002) *Science* 298(5600):1912
- Bamborough P, Brown MJ, Christopher JA, Chung CW, Mellor GW (2011) *J Med Chem* 54(14):5131
- Bhagwat SS (2009) *Curr Opin Investig Drugs* 10(12):1266
- Brandvold KR, Soellner MB (2011) 242nd National meeting of the American-Chemical-Society (ACS), Denver, CO, Aug 28–Sep 01, 2011. Abstracts of papers of the American Chemical Society 242, 338-MEDI
- Cherry M, Williams DH (2004) *Curr Med Chem* 11(6):663
- Daub H, Godl K, Brehmer D, Klebl B, Muller G (2004) *Assay Drug Dev Technol* 2(2):215
- Anastassiadis T, Deacon SW, Devarajan K, Ma HC, Peterson JR (2011) *Nat Biotechnol* 29(11):1039
- Godl K, Daub H (2004) *Cell Cycle* 3(4):393
- Karaman MW, Herrgard S, Treiber DK, Gallant P, Atteridge CE, Campbell BT, Chan KW, Ciceri P, Davis MI, Edeen PT, Faraoni R, Floyd M, Hunt JP, Lockhart DJ, Milanov ZV, Morrison MJ, Pallares G, Patel HK, Pritchard S, Wodicka LM, Zarrinkar PP (2008) *Nat Biotechnol* 26(1):127
- Morphy R (2010) *J Med Chem* 53(4):1413
- Subramanian G, Sud M (2010) *Acs Medicinal Chem Lett* 1(8):395
- Agrawal R, Imielinski T, Swami A (1993) Mining association rules between sets of items in large databases. In: *Proceedings of the 1993 ACM SIGMOD international conference on management of data*. ACM, Washington, DC, USA, p 207
- Brin S, Motwani R, Silverstein C (1997) Beyond market baskets: generalizing association rules to correlations. In: *Proceedings of the 1997 ACM SIGMOD international conference on management of data*. ACM, Tucson, AZ, USA, p 265
- Silverstein C, Brin S, Motwani R (1998) *Data Min Knowl Disc* 2(1):39
- Kinase SARfari. <https://www.ebi.ac.uk/chembl/sarfari/kinasesarfari>. Accessed 2011
- Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP (2012) *Nucleic Acids Res* 40(D1):D1100–D1107. doi:10.1093/nar/gkr777
- Overington JP (2009) 238th National meeting of the American Chemical Society, DC, August 16–20, 2009. Abstracts of papers of the American Chemical Society 238, 39-COMP
- Wadler S (2001) *Drug Resist Updat* 4(6):347
- Bradham C, McClay DR (2006) *Cell Cycle* 5(8):824
- Raymond E, Faivre S, Armand JP (2000) *Drugs* 60(Suppl 1):15
- Chen T, George JA, Taylor CC (2006) *Anticancer Drugs* 17(2):123
- Heron-Milhavet L, Khouya N, Fernandez A, Lamb NJ (2011) *Histol Histopathol* 26(5):651
- Kawakami T, Kawakami Y, Kitauro J (2002) *J Biochem* 132(5):677
- Liew CY, Ma XH, Yap CW (2010) *J Comput Aided Mol Des* 24(2):131
- Han LY, Ma XH, Lin HH, Jia J, Zhu F, Xue Y, Li ZR, Cao ZW, Ji ZL, Chen YZ (2008) *J Mol Graph Model* 26(8):1276
- Yap CW (2011) *J Comput Chem* 32(7):1466
- Durant JL, Leland BA, Henry DR, Nourse JG (2002) *J Chem Inf Comput Sci* 42(6):1273
- Li QL, Chen TJ, Wang YL, Bryant SH (2010) *Drug Discovery Today* 15(23–24):1052
- Bryant S (2006) 231st National meeting of the American Chemical Society, Atlanta, GA March 26–30, 2006. Abstracts of papers of the American Chemical Society 231, 80-COMP
- PubChem Fingerprints. [ftp://ftp.ncbi.nih.gov/pubchem/data\\_spec/pubchem\\_fingerprints.txt](ftp://ftp.ncbi.nih.gov/pubchem/data_spec/pubchem_fingerprints.txt). Accessed 2011
- Klekota J, Roth FP (2008) *Bioinformatics* 24(21):2518
- Japkowicz N, Shah M (2011) *Performance measures I. Evaluating learning algorithms: a classification perspective*. Cambridge University Press, Cambridge
- Rogers DJ, Tanimoto TT (1960) *Science* 132(3434):1115