Reverse Fingerprinting and Mutual Information-Based Activity Labeling and Scoring (MIBALS)

Chris Williams[*,1] and Suzanne K. Schreyer[2]

[1]*Chemical Computing Group Inc., 1010 Sherbrooke Street West, Suite 910, Montreal, Quebec, H3A 27R, Canada*

[2]*Polychromix, 30 Upton Drive, Wilmington, MA 01887, USA*

Abstract: A mutual information based activity labeling and scoring (MIBALS) approach to reverse fingerprint analysis is presented. Whole molecule scores produced by the method are shown to be capable of ranking compounds in virtual high-throughput screening (vHTS) experiments, while fragment scores produced by the method are able to identify pharmacophore moieties important for biological activity. The performance of MIBALS in vHTS experiments is assessed using reference ligands active against 40 different biological targets, and MIBALS retrieval rates are compared with those obtained using more traditional group fusion similarity search methods. The use of MIBALS to identify important pharmacophore fragments is demonstrated by comparing ligand fragment scores with known pharmacophores and known ligand/protein contacts. The ability of MIBALS to highlight beneficial and detrimental groups in a congeneric series is examined by comparing MIBALS fragment scores with features in known structure-activity relationships.

Keywords: Fragment scoring, fingerprints, pharmacophores, similarity searching, virtual screening.

INTRODUCTION

A molecular *fingerprint* is a collection of bits that indicate the presence or absence of certain structural features [1]. The bits themselves may be a host of constructs which include functional groups and chemical fragments [2], extended connectivities [3, 4], descriptor distributions [5] and typed atom polygons [6-9] (Fig. **1**). Fingerprints are typically used in similarity searching [10], compound library clustering [11] and diversity analysis [12].

An attractive feature of many molecular fingerprinting systems is the ease with which an individual bit can be traced back to the atoms and molecular fragments that defined it. Thus, any statistical analyses performed on the fingerprint bits can be mapped back to the molecular fragments used in their construction, thereby producing fragment scores for a given structure as shown in Fig. (**2**).

Indeed, many reports over the past two decades describe pharmacological fragment scoring based on statistical analysis of fingerprint bits. For example, the Stigmata approach used a form of bit frequency to determine structural commonalities within compound datasets [13]. Xue *et al.* have used statistical distributions to isolate bits which perform optimally, for isolating compounds active against different biological targets [14]. Recently, Schneider *et al.* used support vector machines (SVM) to extract and visualize pharmacophore points based on 2D typed-atom triangle fingerprints [15]. For convenience, the term *reverse fingerprinting* (inspired by *reverse QSAR* [16]) will be employed to describe any method that invokes fingerprint bit analysis to arrive at atom or fragment scores. It should be noted that with some fingerprinting systems, such as the Daylight fingerprints [17], the fingerprint is constructed with a coding procedure that loses the correspondence between the fingerprint bits and the molecular fragments; the reverse fingerprinting method described below does not apply to such fingerprinting systems.

In this paper, an information theoretic [18] approach to reverse fingerprinting is presented. The goal is to create a fingerprint-based model that produces whole molecule scores for virtual screening and atom scores for pharmacophore elucidation.

THEORY: APPLICATION OF MUTUAL INFORMATION TO FINGERPRINT BIT SCORING

Applications of mutual information (sometimes called *cross-entropy* [19] in engineering), were initially formulated by Shannon and Weaver [20] in their seminal work in communications theory. Mutual information theory is central to the principal of minimum discrimination information (MDI), which states that given new information, a new distribution can be chosen as similar to the original distribution as possible. Originally formulated to include binary data, it has since been expanded to include continuous variables.

More recently, applications of mutual information theory in computational chemistry include QSAR feature selection [21], generation of topological information indices [22], characterization of partitioned property spaces [23], and the selection of variables for spectral data modeling [24]. Applications in bioinformatics include methodologies to group proteins into appropriate functional families and to determine the similarity of gene and protein sequences [25]. Mutual information applications are then naturally extendable for exploration of fingerprint analyses.

A fingerprint *FP* for molecule *Q* typically contains a number of bits, *k*, which are either 1 or 0 indicating the presence or absence of structural features. Fingerprint bits which

[*]Address correspondence to this author at the Chemical Computing Group Inc., 1010 Sherbrooke Street West, Suite 910, Montreal, Quebec, H3A 27R, Canadá; Tel: +(514)-393-1055, Ext. 50; Fax: +(514)-874-9538; E-mail: cw@chemcomp.com

1386-2073/09 $55.00+.00

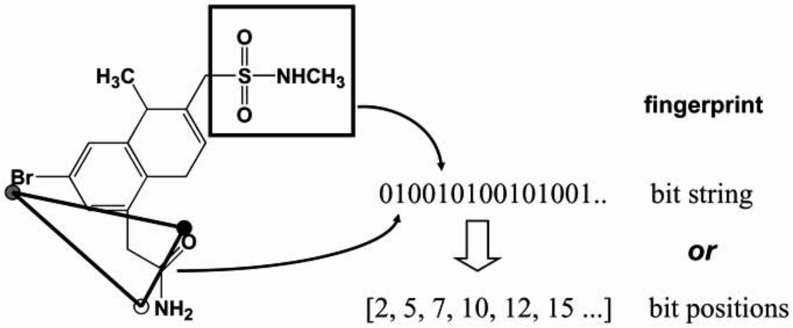


Fig. (1). ***Molecular fingerprints.*** Each bit in the fingerprint indicates the presence/absence (1/0) of a structural motif. The fingerprint may be a *bit string* indicating the on/off state of all the bits.

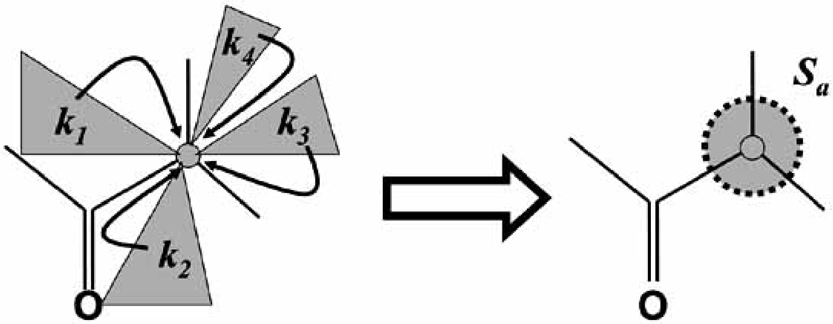


Fig. (2). ***Reverse fingerprinting:*** Atom scores $\mathbf{S_A}$ can be produced by mapping results from statistical analysis of the $\{\boldsymbol{k}_x\}$ fingerprint bits back on to the atoms used to construct the bit.

are present in a given molecule may be grouped together in a fingerprint vector, FP_Q,

$$FP_Q = [k_1, k_2, k_3, \ldots k_n] \tag{1}$$

where k_x is k^{th} fingerprint bit which is turned on in molecule Q. Each bit k in the molecule potentially contains information about the activity state (A) of molecule Q. The *mutual information*, I, between the active state of Q and the k^{th} bit is given by

$$I(A,k) = \sum_{A,k} P(A,k) \log_2 \frac{P(A,k)}{P(A)P(k)} \tag{2}$$

where $P(A)$ and $P(k)$ are the marginal probability distributions of state A and bit k respectively, and $P(A,k)$ is the joint probability distribution of state A and bit k. The sum in equation (2) is over all possible states of A and k, so that mutual information naturally allows for multiple activity states (active, inactive, weakly active, etc.), and multiple bit states (on, off or counts). Note that if the distributions of A and k are mutually independent, the joint probability is equal to the product of the marginal probabilities ($P(A, k) = P(A)$ X $P(k)$) and the mutual information is 0. This additional measure of statistical independence, unlike correlation coefficients, does not require that the variables be normal and linearly distributed.

The mutual information can be conditioned on a set of known actives by rewriting equation (2) as follows:

$$I(A,k) = \sum_{A} \sum_{k} P(A)P(k\,|\,A) \log_2 \frac{P(k\,|\,A)}{P(k)} \tag{3}$$

Confining the activity to two states (A=1, active; A=0, inactive) and the bits to two states (k = 1, bit k is ON; k = 0, bit k is OFF), produces the following four terms from equation (3):

$$\begin{aligned} I(A,k) \;=\;& P(A=1)P(k=1\,|\,A=1)\log_2 \frac{P(k=1\,|\,A=1)}{P(k=1)} \\ +\;& P(A=0)P(k=1\,|\,A=0)\log_2 \frac{P(k=1\,|\,A=0)}{P(k=1)} \\ +\;& P(A=1)P(k=0\,|\,A=1)\log_2 \frac{P(k=0\,|\,A=1)}{P(k=0)} \\ +\;& P(A=0)P(k=0\,|\,A=0)\log_2 \frac{P(k=0\,|\,A=0)}{P(k=0)} \end{aligned} \tag{4}$$

Given a collection of m_1 active compounds and m_0 inactive compounds, the conditional probability for seeing the k^{th} bit in an active compound can be estimated as simply the frequency of the k^{th} bit in the collection of active compounds $\{\mathbf{C}\}$.

$$P(k=1\,|\,A=1) \;=\; C_k \approx \frac{1}{m_1} \sum_{i=1}^{m_1} f_k^i \tag{5}$$

In the above equation f_k^i is either 1 or 0, indicating the presence or absence of the k^{th} bit in the i^{th} molecule. The probability of not seeing the k^{th} bit in an active compound, $P(k=0|A=1)$, is then given by $(1-C_k)$. The conditional probability of the k^{th} bit being on in an inactive compound can be estimated from the set of inactive compounds $\{\mathbf{F}\}$ in a similar manner.

$$P(k=1\,|\,A=0) \;= F_k \approx \frac{1}{m_0}\sum_{i=1}^{m_0} f_k^i \tag{6}$$

The probability of bit k not being turned on in an active compound is then given by (1-F_k). The total probability of seeing bit k in chemical space, $P_k = P(k{=}1)$ can be trained using a dataset {**P**} obtained either from an independent database representing chemical space, or by combining the active and inactive subsets ({**P**} = {**C**} ∪ {**F**}).

The $P(A{=}x)$ terms in equation (4) represent the probability of finding a compound with activity state x. If the probability of finding an active compound is α ($P(A{=}1) = \alpha$) then the probability of finding an inactive compound, $P(A{=}0)$, is given by (1- α). The parameter α can be computed from the training dataset, or chosen depending upon the intended application. Using these prior probability expressions and inserting the expressions for C_k, F_k and P_k into equation (4) yields the following expressions for mutual information contributions:

$$\begin{aligned}
I_{11} = \quad & I(A=1,k=1) &= \alpha\, C_k \log_2 \frac{C_k}{P_k} \\
I_{01} = \quad & I(A=0,k=1) &= (1-\alpha) F_k \log_2 \frac{F_k}{P_k} \\
I_{10} = \quad & I(A=1,k=0) &= \alpha\,(1-C_k)\log_2 \frac{(1-C_k)}{(1-P_k)} \\
I_{00} = \quad & I(A=0,k=0) &= (1-\alpha)(1-F_k)\log_2 \frac{(1-F_k)}{(1-P_k)}
\end{aligned} \tag{7}$$

Summing all four terms in equation (7) will yield the total mutual information between the activity state and the bit state.

To develop an activity score for bit k the individual mutual information terms in equation (7) must be combined to reflect the difference in information about activity and inactivity. The I_{11} and I_{00} terms contribute *affirmative* information about activity - i.e., the compound is active when bit k is present and inactive is bit k is absent. In contrast, the I_{01} and I_{10} terms contribute *dismissive* information about activity - i.e., the compound if inactive when bit k is present and active if bit k is absent. Subtracting the dismissive terms from the affirmative terms leads to the following activity scoring function, S_k, for the k^{th} bit.

$$S_k \;\;=\;\; \alpha(I_{11}+I_{00}) - (1-\alpha)(I_{01}+I_{10}) \tag{8}$$

If we take a naïve-Bayesian approach and assume that the fingerprint bit k distributions are mutually independent (they are not), a score for the entire molecule, S_{M}, can be computed by summing the bit scores over all the unique bits in the query molecule.

$$S_M = \sum_{k}^{\substack{\textit{unique}\\ \textit{bits}}} S_k \tag{9}$$

Although this approximation is somewhat severe, the naïve-Bayesian approach has been successfully applied to fingerprint analysis by a number of workers [26-28]. The molecule score S_{M} can be used for ranking molecules in a vHTS context, or it may be further decomposed into contributions per atom,

$$S_M = \sum_{k}^{\substack{\textit{unique}\\ \textit{bits}}} S_k = \sum_a S_a \tag{10}$$

In equation (10) S_a is the atom score for the a^{th} atom. This score is derived from the S_k scores of the bits in which the atom participates. In this study, only the on/off state (and not the number of instances) of bit k is considered in the calculation of S_k, so the S_k score for a given bit must be equally divided between all n_k instances of than bit. Furthermore, the S_k score for the k^{th} bit must be divided equally between all the atoms that contribute to the construction of the bit - n_k^a. Hence, the S_k score for each bit in which atom a participates must be divided by the instances of the bit in the molecule (n_k) and by the number of atoms contributing to the bit (n_k^{th}). With these considerations, the atom score S_a is given by:

$$S_a = \sum_{k\in a} \frac{S_k}{n_k \times n_a^k} \tag{11}$$

Composed this way, the sum of the atom scores equals the molecule score S_M. Thus, given a fingerprint system, and a set of active and inactive compounds, the C_k, F_k and P_k terms in equation (7) can be trained and used to compute whole molecule scores (S_M) for vHTS, and individual atom scores (S_a) which sum to the molecule score S_M.

The above equations are the basis for training and evaluating mutual information based activity labeling and scoring, or *MIBALS* models. The following procedures outline the methods for fitting, evaluating and producing atom scores from MIBALS models.

MIBALS Model Fitting Procedure

1. Select {**C**}, {**F**} and {**P**} training sets.
2. Fit $\mathbf{C}_{\boldsymbol{k}}$, $\mathbf{F}_{\boldsymbol{k}}$ and $\mathbf{P}_{\boldsymbol{k}}$, and compute $\mathbf{S}_{\boldsymbol{k}}$ for all unique $\boldsymbol{k}$ bits.
3. Normalizing Factor: ***Z***
   (i) compute the atom scores for all compounds in {**C**}
   (ii) determine ***Z***; ***Z*** = 100 / max { **C** atom scores }
4. Return ***k***, $\mathbf{S}_{\boldsymbol{k}}$, and ***Z***.

MIBALS Model Evaluation Procedure

1. For a molecule ***M*** with bits $\boldsymbol{k_m}$, select $\boldsymbol{k_m} \in \boldsymbol{k}$.
2. Compute the molecule score $\boldsymbol{S_M} = Z \times \sum_{k\in k_m}^{m} S_k$

MIBALS Atom/Fragment Scoring and Visualization

1. Given a query molecule ***M*** with bits $\boldsymbol{k_m}$, select $\boldsymbol{k_m} \in \boldsymbol{k}$.
2. For each atom/fragment, evaluate $\boldsymbol{S_a}$ using equation (11).
3. Visualize the fragment scores by drawing spheres centered on the atom fragments. The sphere radii $\boldsymbol{r_a}$ are scaled to reflect the magnitude of $\boldsymbol{S_a}$

   $\boldsymbol{r_a}$ (Å) = $\boldsymbol{S_a}$ / 100.
4. Render the spheres to reflect the sign of $\boldsymbol{S_a}$; positive $\boldsymbol{S_a}$,=solid, negative $\boldsymbol{S_a}$=line.

Selection of the {**C**}, {**F**} and {**P**} training sets in the MIBALS model fitting procedure will depend on the intended use of the model. Since the goal of vHTS work is to separate potentially active compounds and chemotypes from random chemical space, MIBALS models built for vHTS purposes would typically use a set of active compounds for the {**C**} set and a diverse sample of drug-like molecules for the {**P**} set. Since there are no *focused* inactives (inactive compounds with the same chemotype as the actives) in this context, the {**F**} set would be empty ({**F**} =∅). The MIBALS models thus generated would be applied to ranking molecules in a vHTS context – separating active chemotypes from random chemical space. The atom/fragment scores generated by these models would isolate the core chemical features that most distinguish active chemotypes from random chemical space. The fragment scores may also reflect the pharmacophore for the active chemotype(s).

MIBALS models can also be applied to the analysis of a congeneric series, helping to determine fragments that diminish and/or enhance activity with respect to the active core of the chemotype. In this case, the {**C**} set would be composed of active members of the congeneric series, while the {**F**} set would consist of inactive congeners. The {**P**} set could be composed of either a union of the {**C**} and {**F**} sets, or an independent diverse set of random drug-like molecules.

To investigate the behavior of the MIBALS models with respect to vHTS experiments, pharmacophore elucidation and congeneric series analysis, the following studies were performed:

1. <u>*MIBALS Models: General vHTS Experiment*</u> – 40 biological targets were chosen for vHTS experiments with mutual information models. Details of the data source, a list of the biological targets and the number of active ligands per target are given in the **Methods** section. For each biological target, MIBALS models were made using reference sets {**C**} of various sizes, consisting of ligands active against the target in question. The {**F**} set was empty in all cases, ({**F**} =∅) and the {**P**} set consisted of 1600 diverse drug-like molecules. The MIBALS model retrieval rates were compared to retrieval rates obtained using multiple reference compound similarity searches using the *maximum* (MAX) and *average* (AVE) data fusion rules as described by Ginn *et al.* [29]. The data fusion rules produce similarity scores by applying the following conditions to the set of similarities between the test molecule *t*, and all of the reference molecules $\{Q_i\}$.

   AVE–fusion score is an *average* of the similarities $\boldsymbol{S} = \text{ave}\ \{S(Q_i,t)\}$

   MAX –fusion score is the *maximum* of the similarities $\boldsymbol{S} = \max\ \{S(Q_i,t)\}$

   where $\{S(Q_i,t)\}$ is the set of similarities between all of the reference group molecules Q_i and the test molecule *t*. To facilitate discussion the number of compounds in the reference group will be referred to

Table 1. Biological Target Codes: The Names and Abbreviations (Codes) of the 40 Biological Targets Considered in the vHTS Study

Target	Code	Target	Code
5-HT1A serotonin	5-HT1A	Endothelin A	ETA
5-HT1B serotonin	5-HT1B	Factor II alpha	FIIa
5-HT2A serotonin	5-HT2A	Farnesyl tranferase	FTase
5-HT2C serotonin	5-HT2C	Factor X alpha	FXa
5-HT4 serotonin	5-HT4	HIV reverse transcriptase	HIV-1-RT
Acetylcholinesterase	AChE	HIV protease	HIVPR
Adenosine A1 receptor	ADORA1	Hydrolase (multiple targets)	HLase
Alpha-1 adrenergic receptor	ADRA1	Histamine Receptor H3	HRH3
Adenosine kinase	AK	Kappa opioid receptor	KOR
Butylcholinesterase	BChE	Matrix metalloprotease-3	MMP-3
Carbonic anhydrase 1	CA-1	Matrix metalloprotease-8	MMP-8
Carbonic anhydrase 2	CA-2	Micro-opioid receptor	MOR
Cannabinoid CB1	CCB1	Melatonin MT1 receptor	MT1
Cholecystokinin B	CCKB	Neuropeptide Y receptor 5	NPY5
CDK2/Cyclin-A-dependant Kinase	CDK2-Cyclin_A	cAMP phosphodiesterase 4	PDE4
Cyclooxygenase-2	COX-2	cGMP phosphodiesterase 5	PDE5
Dopamine D2	D2	Protein tyrosine phosphatase 1B	PTP-1B
Dihydrofolate reductase	DHFR	Steroid receptor coactivator	SRC
Dimerization partner 2	DP2	Tumor necrosis factor-alpha converting enzyme	TACE
Epidermal growth factor receptor	EGFR	Trypsin	Trypsin

as *n*. The MAX rule was chosen for the validation of MIBALS because studies to date suggest the MAX fusion rule is the most effective for fingerprint based similarity searches [30, 31].

2. MIBALS Pharmacophore Elucidation – MIBALS models for PDE4 and FXa were constructed with the same training sets used in the general vHTS experiments ($\{\mathbf{C}\}$ = active ligands, $\{\mathbf{F}\} = \varnothing$, $\{\mathbf{P}\}$ = diverse drug-like molecules). The MIBALS models were used to score atoms in the X-ray determined ligand structures 1R06 and 1FAX. The neutral 1R06 ligand (rolipram) was untouched from the crystal structure. The amidine group of the 1FAX ligand was protonated to produce the amidinium group. Protein residues were assigned standard (pH=7) protonation states. The pattern of fragment scores of both ligands were compared with binding contacts known from the crystal structures.

3. MIBALS Congeneric Series Analysis – a congeneric series of PDE4 inhibitors was chosen for analysis with the MIBALS models. The $\{\mathbf{C}\}$ reference set contained all the *active* congeners (pIC$_{50}$>7) while the $\{\mathbf{F}\}$ set contained the inactive congeners. The $\{\mathbf{P}\}$ set was the entire series ($\{\mathbf{P}\} = \{\mathbf{C}\} \cup \{\mathbf{F}\}$).

METHODS

vHTS Test and Training Datasets: The *entire* dataset used in this study consisted of 10106 unique compounds with biological measurements spanning 438 biological targets, compiled from *Journal of Medicinal Chemistry* articles covering the years 1994 to 2004. The data used is a subset of the ChemBioBase^{TM} database available from Jubilant Biosys Ltd. [32]. Details of the dataset preparation are available upon request. Forty (40) of the biological targets (listed in Table **1**) were chosen to study the vHTS recall rates of MIBALS models.

It is important to note that the data contains many congeneric series for each target, and there are many examples of active and inactive compounds for each target/chemical class combination. Since most of the measurements in the dataset are continuous, consisting mainly of pIC$_{50}$ and IC$_{50}$ values in different units, they were all normalized to pIC$_{50}$ (M) activity values, and a threshold of pIC$_{50}$ (M) > 6 was set to separate the 'active' compounds for the study. The normalization procedure was less than perfect (due to variation in activity experiments among different journal articles), so this procedure introduced some noise into the activity data. In addition, the pIC$_{50}$ threshold of 6 is also somewhat arbitrary, and other thresholds could have been chosen. By computing the number of active compounds at three different activity thresholds (pIC$_{50}$ = 4, 6, 8), one can see that compounds for each of the forty targets span a number of activity ranges (Table **2**).

Table 2. Number of Active Compounds for Each Biological Target at Various pIC$_{50}$ Thresholds

# of Actives at Various pIC$_{50}$ Thresholds							
Code	**pIC$_{50}$ > 4**	**pIC$_{50}$ > 6**	**pIC$_{50}$ > 8**	**Code**	**pIC$_{50}$ > 4**	**pIC$_{50}$ > 6**	**pIC$_{50}$ > 8**
5-HT1A	505	480	165	ETA	417	375	207
5-HT1B	161	118	9	FIIa	376	158	58
5-HT2A	214	193	58	FTase	327	216	55
5-HT2C	145	122	23	FXa	330	221	63
5-HT4	91	89	15	HIV-1-RT	198	121	4
AChE	186	87	30	HIVPR	226	219	186
ADORA1	457	355	55	HLase	304	211	16
ADRA1	573	522	161	HRH3	92	88	34
AK	107	94	36	KOR	320	248	106
BChE	109	70	9	MMP-3	477	418	117
CA-1	696	483	30	MMP-8	313	279	148
CA-2	723	701	290	MOR	381	301	153
CCB1	107	98	27	MT1	170	164	102
CCKB	81	73	21	NPY5	188	158	61
CDK2-Cyclin_A	123	97	42	PDE4	221	162	44
COX-2	244	112	6	PDE5	549	400	234
D2	616	535	116	PTP-1B	210	102	0
DHFR	175	132	47	SRC	344	225	48
DP2	85	84	5	TACE	144	141	34
EGFR	359	184	61	Trypsin	281	107	1

Number of active compounds in the entire database at three different activity thresholds (pIC$_{50}$ = 4, 6, 8) for each of the 40 biological targets considered in the vHTS study. At a pIC$_{50}$ threshold of 4, the total number of active compounds is greater than the entire dataset because many compounds are active against multiple targets.

Choosing an activity threshold of $pIC_{50} > 6$ means that the resulting data set contains a substantial number of inactive compounds which belong to the same chemical classes as actives, and lie on the interface region of pIC_{50}s between 4 and 6. At an activity threshold of $pIC_{50} > 4$, the sum of all the active compounds is *greater* than the number of compounds in the dataset, because many compounds are active against a number of the targets at this threshold, and thus count for more than one active compound when computing the total number of active compounds in the dataset.

A *training set* of 1600 unique compounds was constructed by randomly extracting from the entire 10106 compound dataset 40 active compounds for each of the 40 targets in Table **1**. All of the training compounds were removed from the entire dataset to create an 8506 compound *external test set* which was devoid of any of the training compounds. The external test set was used for determining the recall rates of the MIBALS models.

MIBALS Model Application Development: The scientific vector language (SVL) in the MOE software [33] was used to code up applications for MIBALS model fitting, MIBALS model evaluation and MIBALS based fragment scoring. These applications are available from the SVL program exchange (svl.chemcomp.com) under the title "Reverse Fingerprinting: Mutual Information Based Activity Labeling and Scoring (MIBALS)".

General vHTS Experiment: MIBALS models were made for biological activity against each target in Table **1** by using reference groups of various sizes consisting of compounds active against the target. Two different 2D fingerprinting schemes were used; the fragment-based 166 MDL MACCS key fingerprint [34] and the PCH fingerprint. The PCH fingerprint is a typed graph triangle based fingerprint that uses the pharmacophore annotation points in the MOE software as triangle vertices. Details of the PCH implementation were presented in a previous publication [35]. Reference active compounds sets $\{\mathbf{C}\}$ of the following sizes – (n = 1, 7, 15, 25) - were chosen for each target. In all cases the $\{\mathbf{P}\}$ set was taken to be the entire 1600 compound training set. To compare the MIBALS models with multiple reference similarity searches, the MOE 2005.06 Fingerprint-Model application was used to create MAX and AVE data fusion fingerprint models with the $\{\mathbf{C}\}$ reference sets.

The MIBALS, MAX and AVE models were then used to virtually screen the external test dataset for active compounds. The overall success of each screen was assessed using the area under the receiver operating characteristic (ROC AUC) [36, 37] curve. The scale for the ROC AUC was chosen so that 1.00 corresponds to a perfect model while 0.5 is a random model. In addition to the ROC AUC, the ability of the methods to retrieve compounds early on in the virtual screening runs was assessed using enrichment factors (EF), computed as the percent active compounds retrieved when 1% (EF@1%) and 10% (EF@10%) of the test database has been filtered.

Reference compounds were chosen by randomly selecting an appropriately sized subset (n = 1, 7, 15, 25) from the 40 training compounds active against the target in question. This procedure was repeated five (5) times for each value of n, and the ROC AUC values averaged over all repetitions, to give the ROC AUC value for each reference group size/biological target combination. For each repetition, MIBALS models were made using three different values of α – 0.5, 0.01 and 0.001 - in order to investigate the effect of this parameter on the MIBALS models.

MIBALS Pharmacophore Elucidation: To test MIBALS pharmacophore elucidation, X-ray structures of the biological targets PDE4 and FXa with bound ligands (PDB codes 1R06 and 1FAX respectively) were used as references for 'true' pharmacophores. MIBALS models for PDE4 and FXa activity were constructed using reference sets ($\{\mathbf{C}\}$) of 20 active compounds for each target, randomly selected from the 1600 compound training database. The $\{\mathbf{P}\}$ set consisted of the entire 1600 training database. Both the MACCS and PCH fingerprints were used. MIBALS models were constructed for both systems with the α parameter set to 0.5. These models were then used to score the fragments in the PDB ligands.

MIBALS Congeneric Series Analysis: The ability of MIBALS models to detect beneficial and detrimental groups in a congeneric series was investigated using a set of 50 compounds with known PDE4 activity. The $pIC_{50}(M)$ values for the compounds ranged from 5.2 to 9.2. An activity threshold was set such that compounds with a $pIC_{50} \geq 7.0$ are deemed active. At this threshold the dataset consisted of 32 actives and 18 inactives. These compounds can be divided into two related series – series 1 in Table (**3**) and series 2 in Table (**4**). Each series has examples of active and inactive compounds at an activity threshold of $pIC_{50} \geq 7$.

The $\{\mathbf{C}\}$ set was comprised the active compounds from both series, while the $\{\mathbf{F}\}$ set consisted of the inactive compounds from both series. The $\{\mathbf{P}\}$ set was taken to be the union of $\{\mathbf{C}\}$ and $\{\mathbf{F}\}$ ($\{\mathbf{P}\} = \{\mathbf{C}\} \cup \{\mathbf{F}\}$). MIBALS models were constructed using both the MACCS and PCH fingerprints. Since this case is an example where the number of known actives and inactives is approximately equal, a value of 0.5 was used for α. The resulting model was used to score atom fragments in representative active and inactive congeners shown in Fig. (**3**).

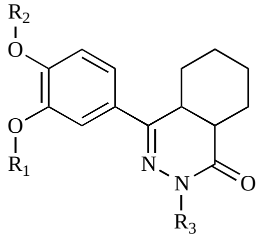
RESULTS AND DISCUSSION

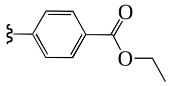
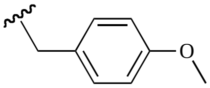
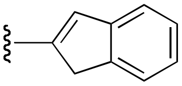
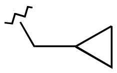
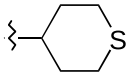
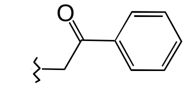
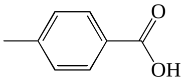
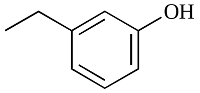
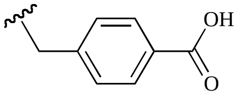
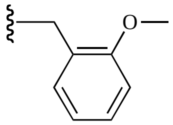
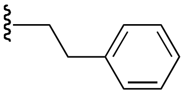
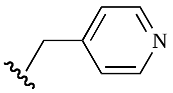
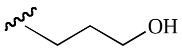
MIBALS Models: General vHTS Experiment

The general effect of reference group size (n) on the recall performance of MIBALS models is demonstrated in Figs. (**4**) and (**5**), which plot the average ROC AUC over all 40 biological targets as a function of n. For comparison, the average ROC AUCs for the MAX and AVE fusion rule similarity searches are also given.

The plots show that both the MACCS and PCH fingerprint MIBALS models, on average, perform at least as well as the AVE fusion rule similarity searches. In contrast, MACCS key MIBALS models generally produce smaller ROC AUC values than MACCS similarity searches using the MAX fusion rule, with the differences between the two increasing as a function of n. MIBALS models made with PCH fingerprint bits produce ROC AUC values comparable to those produced by MAX fusion rule similarity searches up to n = 15, with only a small drop-off in ROC AUC by n = 25. The comparable performance of the MIBALS and AVE fusion rule models over all targets and reference group sizes

Table 3. *PDE4 Congeners - Series 1:* Structures Used in PDE4 Active Congeneric Series Study. pIC_{50}(M) Values are Given for Each Compound

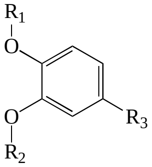


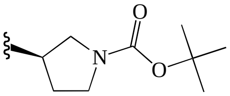
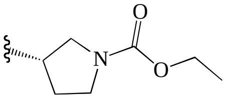
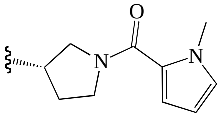
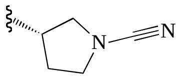
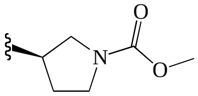
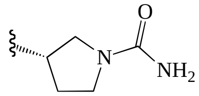
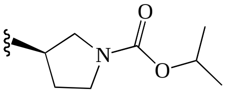
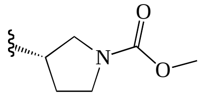
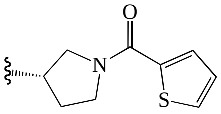
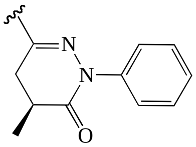
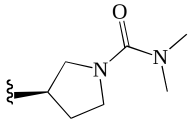
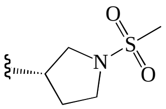
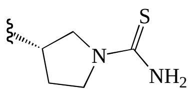
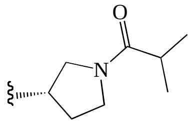
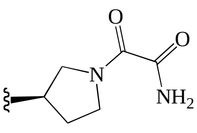
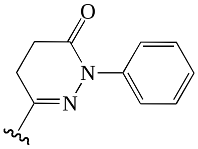
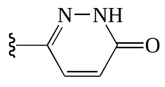
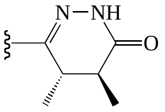
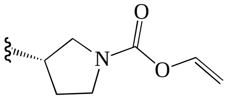
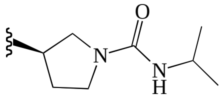
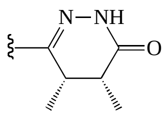
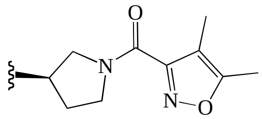
	PDE4 pIC_{50} (M)	R1	R2	R3		PDE4 pIC_{50} (M)	R1	R2	R3
1a	9.20	CH_3	CH_3	adamantyl	**1h**	8.20	ethyl	ethyl	
1b	9.20	CH_3	CH_3	*c*-hexyl	**1i**	8.20	CH_3	CH_3	
1c	9.10	CH_3	CH_3		**1j**	8.10	CH_3	CH_3	
1d	9.10	CH_3	CH_3	*c*-heptyl	**1k**	8.10	CH_3	CH_3	*n*-butyl
1e	9.00	*c*-propyl	CH_3	*c*-heptyl	**1l**	8.00	CH_3	*c*-propyl	*n*-butyl
1f	8.90	CH_3	CH_3		**1m**	8.00	CH_3	CH_3	
1g	8.30	ethyl	ethyl		**1n**	8.00	CH_3	CH_3	
1o	7.90	CH_3	CH_3	*t*-butyl	**1v**	7.40	CH_3	CH_3	$CH_2C{\equiv}CH$
1p	7.80	CH_3	CH_3	Isopropyl	**1w**	7.30	CH_3	CH_3	
1q	7.70	CH_3	CH_3		**1x**	7.10	*c*-pentyl	CH_3	H
1r	7.60	CH_3	CH_3		**1y**	6.80	CH_3	CH_3	CH_3
1s	7.60	CH_3	CH_3		**1z**	6.40	CH_3	CH_3	H
1t	7.50	CH_3	CH_3		**1aa**	6.40	CH_3	CHF_2	H
1u	7.40	CH_3	CH_3	ethyl	**1bb**	6.10	CH_3	CHF_2	phenyl

suggests that the MIBALS molecule scores are reasonable for similarity searching, although not optimal.

The average ROC AUC produced by models trained with 15 reference compounds is plotted for each biological target in Figs. (**6**) and (**7**). For comparison, the ROC AUCs pro-

Table 4. *PDE4 Congeners* - Series 2*:* Structures Used in PDE4 Active Congeneric Series Study

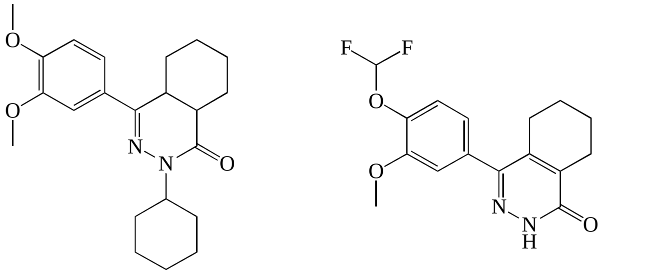


	PDE4 pIC$_{50}$ (M)	R1	R2	R3		PDE4 pIC$_{50}$ (M)	R1	R2	R3
2a	7.52	*c*-pentyl	CH$_3$		**2f**	7.15	*c*-pentyl	CH$_3$	
2b	7.25	*c*-pentyl	CH$_3$		**2g**	7.05	*c*-pentyl	CH$_3$	
2c	7.23	*c*-pentyl	CH$_3$		**2h**	7.00	*c*-pentyl	CH$_3$	
2d	7.19	*c*-pentyl	CH$_3$		**2i**	696	*c*-pentyl	CH$_3$	
2e	7.15	*c*-pentyl	CH$_3$		**2j**	6.90	CH$_3$	CH$_3$	
2k	6.84	*c*-pentyl	CH$_3$		**2q**	6.55	*c*-pentyl	CH3	
2l	6.81	*c*-pentyl	CH$_3$		**2r**	6.50	*c*-pentyl	CH3	
2m	6.80	*c*-pentyl	CH$_3$		**2s**	6.40	CH3	CH3	
2n	6.80	CH3	CHF$_2$		**2t**	6.20	CH3	CH3	
2o	6.71	*c*-pentyl	CH$_3$		**2u**	6.13	*c*-pentyl	CH3	
2p	6.70	CH3	CH3		**2v**	6.10	*c*-pentyl	CH3	

pIC$_{50}$(M) values are given for each compound.

duced from MAX and AVE fusion rule similarity searches are also shown.

The plots in Figs. (**6**) and (**7**) show that the absolute and the relative recall performance is both method and target dependant. For some targets all three model types produce similar ROC AUCs, while for other targets there is wide variation in the recall results produced by each method. For example, the MAX rule exhibits much better recall rates than either the MIBALS or AVE models for some targets (e.g., BChE with the MACCS keys), while for other targets (e.g., COX-2 with MACCS keys), the MIBALS model produces a better recall rate.



1 (PDE4 pIC50: 9.2) **2** (PDE4 pIC50: 5.2)

Fig. (3). ***PDE4 congeneric series analysis*****:** Sample PDE4 active and inactive structures used to demonstrate MIBALS fragment scoring of congeneric compounds.

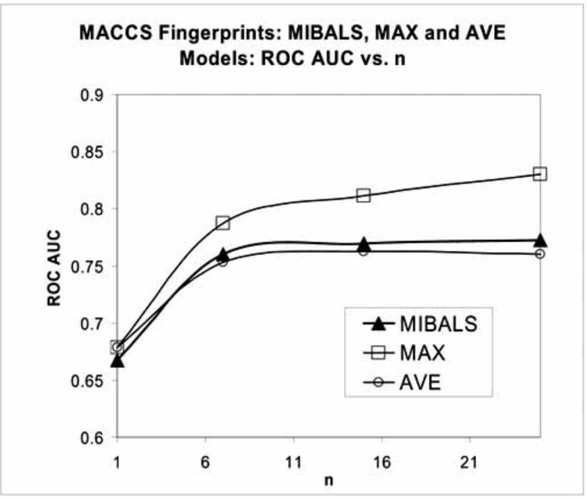


Fig. (4). ***MACCS fingerprint - ROC AUC vs n*****:** The MACCS key ROC AUCs averaged over all 40 biological targets as a function of the number of reference compounds ***n***.

The somewhat poorer recall performance of MIBALS models relative to MAX similarity searching as the reference set size increases is at first glance somewhat discouraging, especially since MAX fusion rule similarity searching is conceptually much simpler than the MIBALS models. However, the differences can be rationalized if one considers the effect of reference set size and diversity on similarity searching results; similarity searching in chemical space can be represented pictorially using similarity score contours - in Fig. (**8**), each point represents a compound, with the distances between the points measuring the *dissimilarity*, *D*, between the compounds (where *D* = 1- *Sab*, and *Sab* is the similarity between compounds *A* and *B*).

The black square in Fig. (**8**) represents a single reference compound *G*, while the grey circles are active compounds that ideally would be retrieved in the virtual screen. When only one reference compound is used, the MAX and AVE fusion rules produce identical scores, which form concentric circles around the reference compound. However, differences in behavior between the MAX and AVE fusion rules become apparent as the size and diversity of the reference set increases. In Fig. (**9**), the AVE and MAX fusion rule similarity score contours are drawn for three different reference set scenarios.

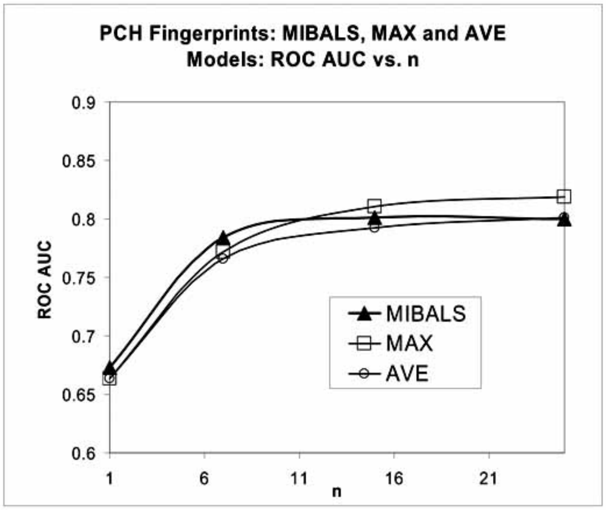


Fig. (5). ***PCH fingerprint - ROC AUC vs n*****:** The PCH fingerprint ROC AUCs averaged over all 40 biological targets as a function of the number of reference compounds ***n***.

When the reference compounds are clustered closely in chemical space as shown in Fig. (**9a**), the AVE and MAX fusion rules produce similar score isocontours, and comparable recall rates would be expected from similarity searches using either rule. However, as the reference compounds become increasingly diverse, as shown in Fig. (**9b**) and (**9c**), the AVE rule isocontours center on the average chemical space of the reference compounds, while the MAX rule contours outline the different clusters of reference compounds. Thus, different recall rates are expected from the AVE and MAX fusion rules when diverse reference sets are used for training.

The previous discussion suggests that MIBALS models would produce results similar to the AVE fusion rule because both approaches produce scores derived from all the reference compounds, as opposed to the MAX fusion rule scores, which reflect only the reference compound most similar to the query. Further discrepancy is expected between MIBALS scores and MAX and AVE similarity scores because MIBALS models lump the fingerprint bits from *all* the reference compounds into one score, without regard to how the bits were grouped in the original compounds. This is one effect of the assumption of bit independence made in equation (9) - information about bit correlations contained in the original reference compounds (i.e., active compounds have bit A or B, but not both) is lost in the MIBALS score. Detrimental effects due to bit correlation are expected to be more prominent when using fingerprints with a small number of unique bits such as the MACCS keys, because there are fewer bits to describe the entirety of chemical space, and thus there is a greater likelihood that active and inactive

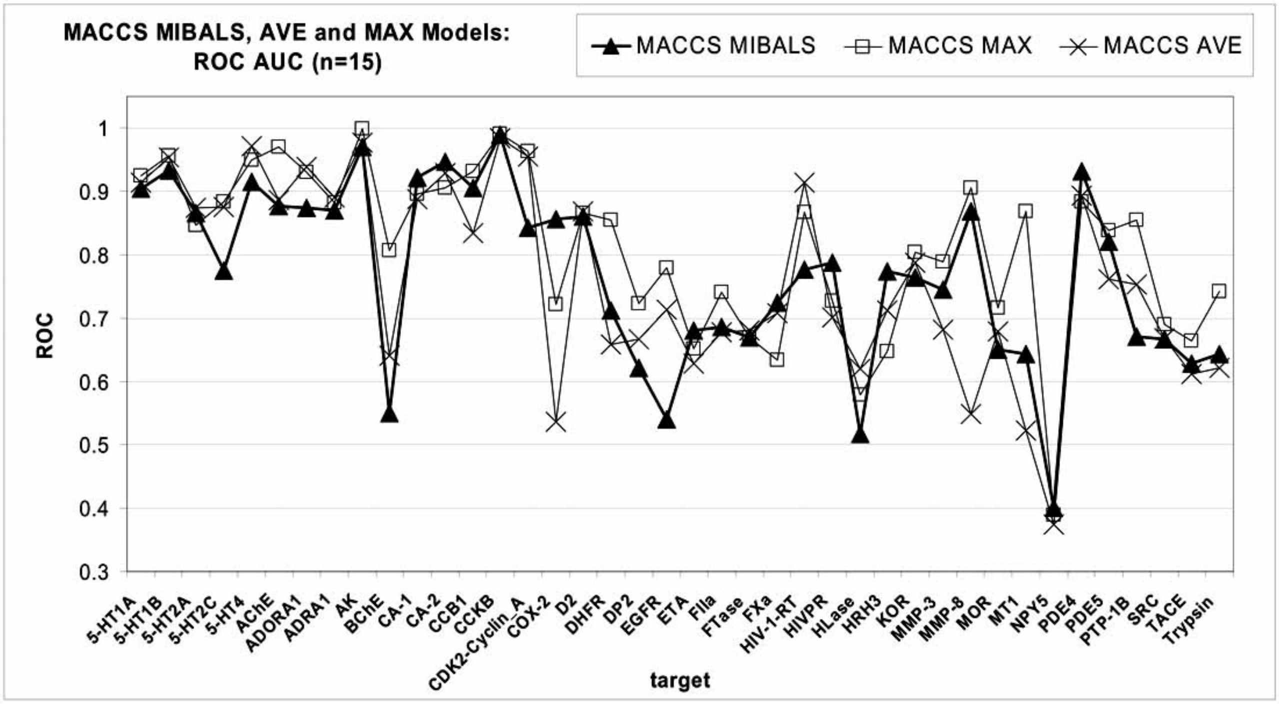


Fig. (6). ***MACCS fingerprint:*** Average ROC AUC with *n=15* reference structures. MIBALS, MAX and AVE models.

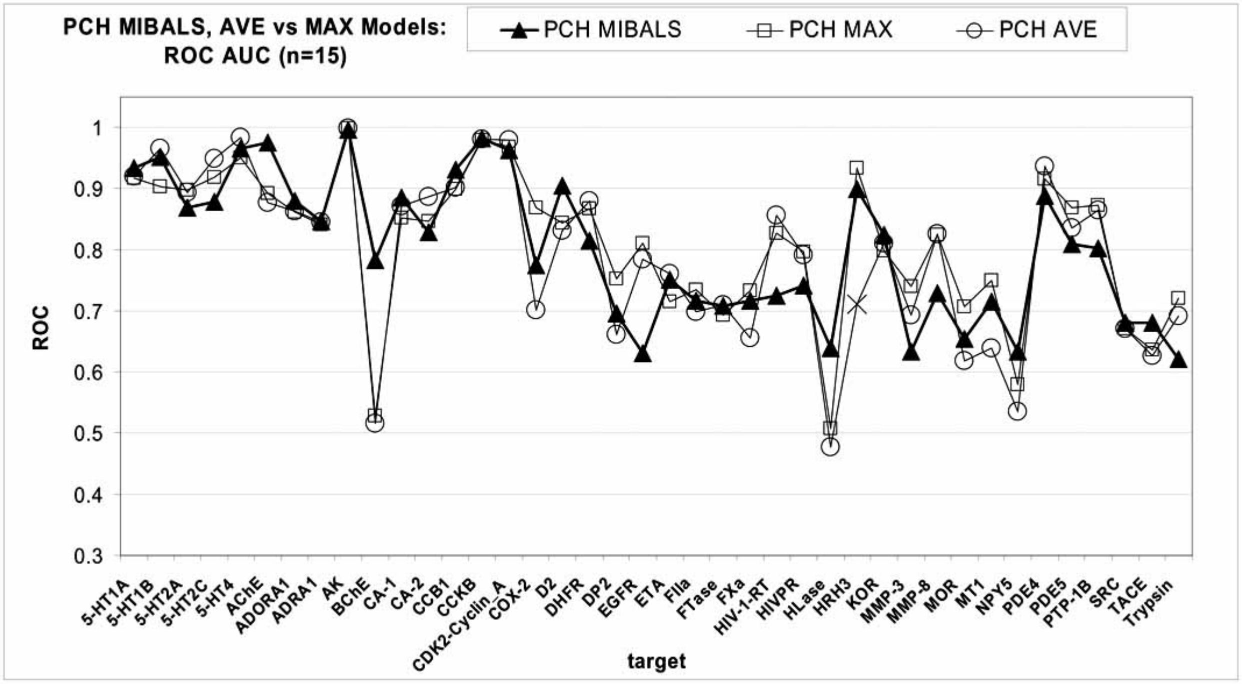


Fig. (7). ***PCH fingerprint:*** Average ROC AUC with *n=15* reference structures. MIBALS, MAX and AVE models.

compounds share the same bits. Indeed, MIBALS models made with the MACCS keys always under-perform the MAX similarity search results, while MIBALS models made with the much larger key set of the PCH fingerprint have comparable perform to MAX similarity searching.

Further insight into the relative behavior of the MIBALS and similarity searching models can be gained from considering the enrichment factors as a function of reference group size. In Figs. (**10**) and (**11**) the enrichment factors at 1% and 10% of the database filtered are given as a function of

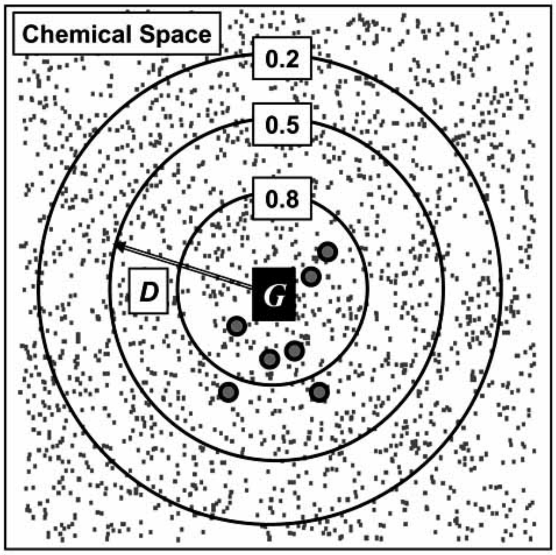


Fig. (8). Similarity isocontours (single reference structure): Plot points represent compounds in chemical space and distances between points measure *dissimilarity* (*D*), given by *D* = 1- *Sab*. The black square represents a single reference compound *G*, and the grey circles are active compounds that ideally will be retrieved in the virtual screen. With a single reference compound, the *similarity score*, S_{AB} is simply the similarity of the test compound with *G* – these are plotted at approximate isocontours of S_{AB} = 0.8, 0.5 and 0.2. The similarity score isocontours form concentric circles in chemical space centered around the reference compound.

reference group size (*n*) for the data fusion similarity models and for the MIBALS models. The plots in Fig. (**10**) show that both at the 1% and the 10% database filtered levels, the MACCS MIBALS enrichment factors as a function of *n* closely follow those of the AVE fusion rule enrichment factors, and except for the case of *n*=1, are lower then the MAX rule enrichment factors. The plot in Fig. (**11**) shows that both at the 1% and the 10% database filtered levels, the PCH MIBALS enrichment factors more closely follow the PCH MAX fusion rule enrichment factors as a function of *n*, and are always greater than the AVE fusion rule enrichment factors. Furthermore, up to a reference group size of *n*=7, the PCH MIBALS enrichment factors are larger than the PCH MAX fusion rule enrichment factors, reflecting the results in Fig. (**5**) which show that up to *n*=7, the PCH MIBALS models yield on average slightly higher ROC AUC values that the PCH MAX fusion rule models. Interestingly, the MIBALS enrichment factors seem less sensitive to changes in *n* than the MAX of AVE fusion rule enrichment factors.

The effect of different α parameters on the MIBALS models is demonstrated in Figs. (**12**) and (**13**). Here, the average ROC AUC for each biological target obtained with 15 reference compound (*n* = 15) MACCS and PCH MIBALS models is plotted for three different values of α; 0.001, 0.01 and 0.5.

The plots in Figs. (**12**) and (**13**) show that the α parameter has little effect on the recall rates of the methods. This may partially be expected, because in the vHTS experiments performed here the 'inactive' compounds are random drug-like molecules, with near-zero information in their bits. Thus,

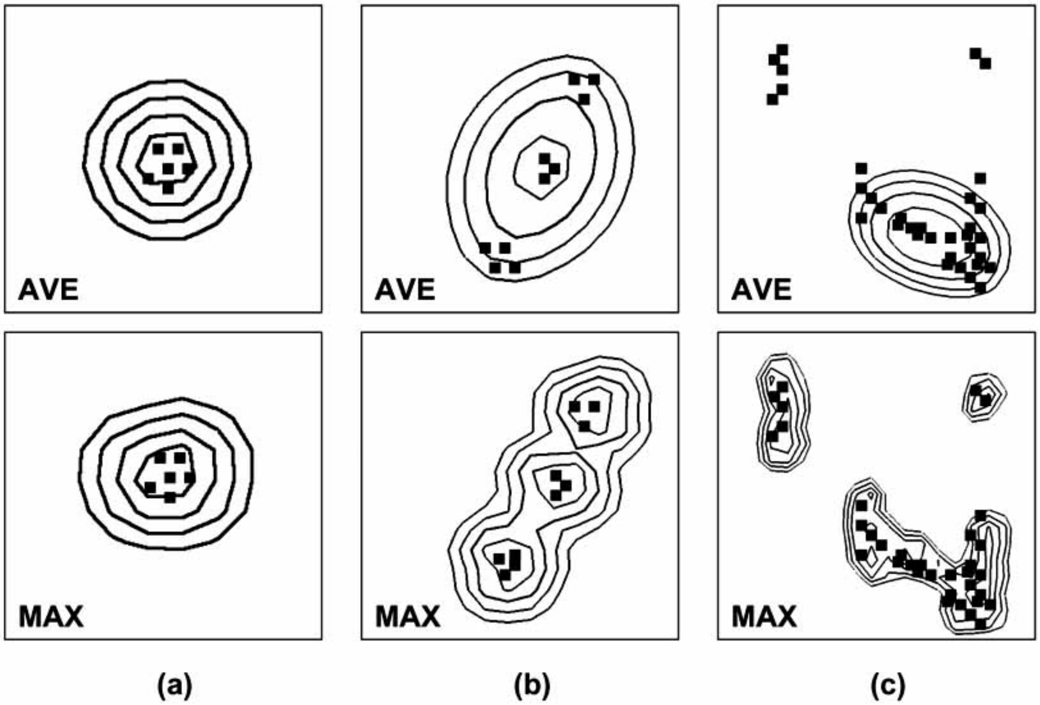


Fig. (9). AVE and MAX similarity score isocontours (multiple reference structures): (a) When reference structures all cluster together tightly in chemical space, the MAX and AVE fusion score isocontours are similar. **(b)** When reference structures form tight but proximal clusters, the MAX and AVE score contours begin to differ. **(c)** When reference structures spread unevenly in chemical space, the MAX and AVE score contours are quite dissimilar, with the MAX score more closely capturing the known actives.

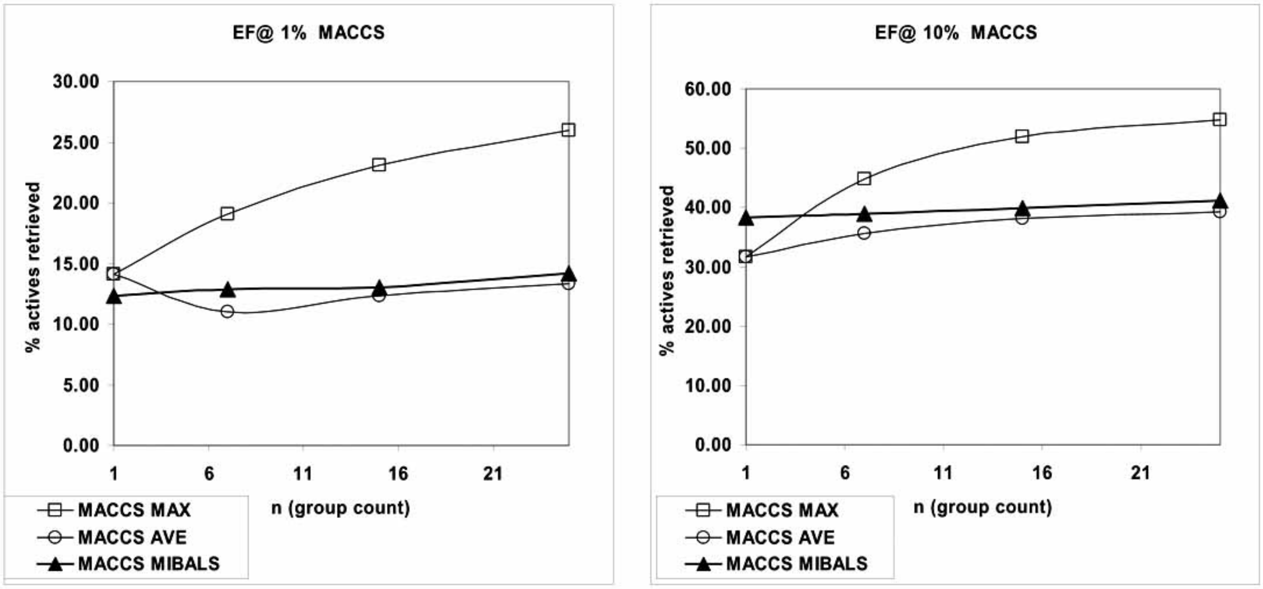


Fig. (10). ER curves for MACCS: EF@1% and EF@10% *vs n* for MAX, AVE and MIBALS models.

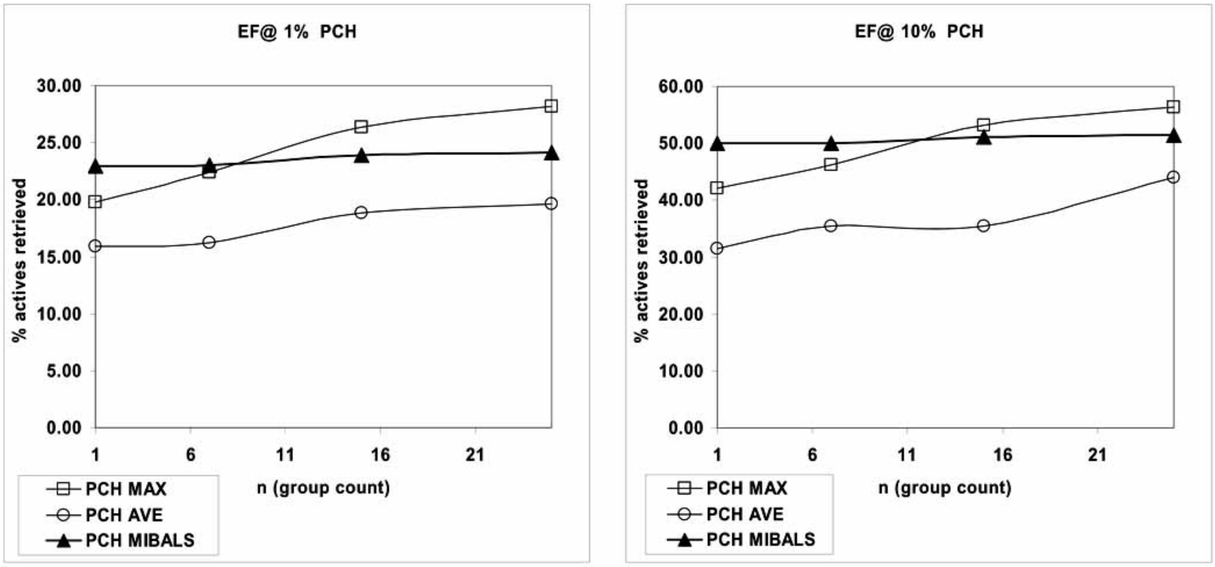


Fig. (11). ER curves for PCH models. EF@1% and EF@10% *vs n* for MAX, AVE and MIBALS models.

even though (1- α) > α at small values of α, the bit score S_k is still dominated by the I_{11} and I_{00} terms, because the information in the I_{01} and I_{10} terms is expected to be near zero. Experiments to investigate the effect of increasing the importance of the I_{01} and I_{10} terms by adding focused inactives to the training set will be the subject of future work.

MIBALS Pharmacophore Elucidation

The MACCS and PCH fingerprint MIBALS atom scores for DX-9065 are drawn in Figs. (**14**) and (**15**), along with a 2D representation of the 1FAX active site residues. H-bond interactions between the ligand and the receptor are shown as solid arrows. The fragment scores are also shown. For clarity, the fragment scores have been rounded to the nearest integer, and only scores with absolute values $\geq$ 10 are shown.

Both the MACCS and PCH fingerprints produce atom scores for DX-9065 that reflect the important protein-ligand interactions. The highest scoring atoms are those that either directly form H-bonds with the receptor, or are in fragments that H-bond with the receptor.

Similarly, the MACCS and PCH fingerprint MIBALS atom scores for rolipram are drawn in Figs. (**12**) and (**13**), along with a 2D representation of the 1R06 active site.

Both the MACCS and PCH fingerprints produce rolipram atom scores that reflect the important protein-ligand interac-

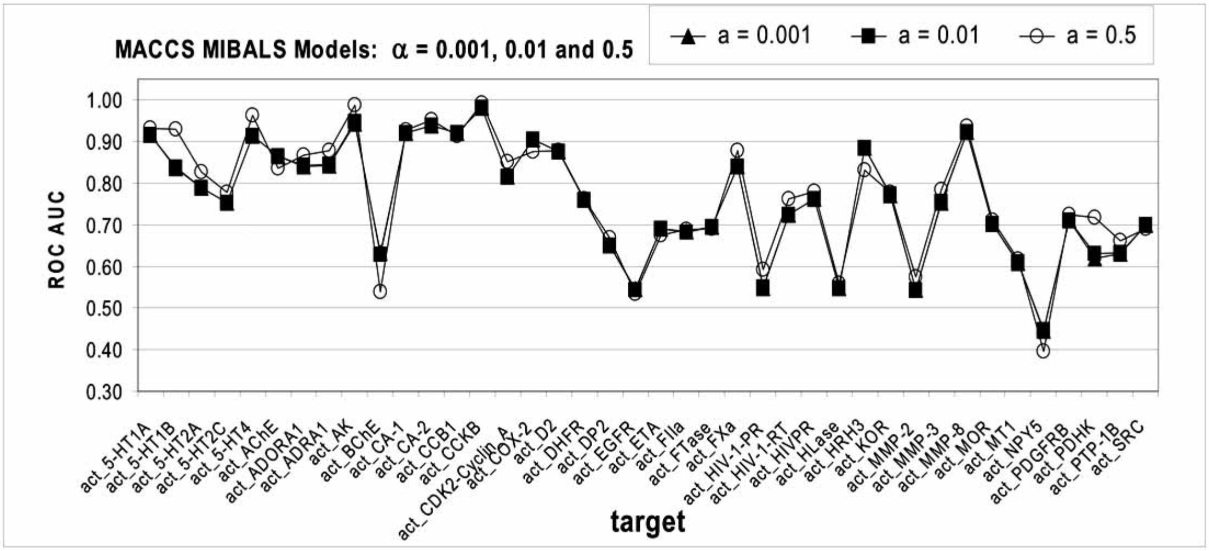


Fig. (12). MACCS MIBALS models: Average ROC AUC v. target (*n=15* reference structures) - effect of α (α = 0.001, 0.01, 0.5).

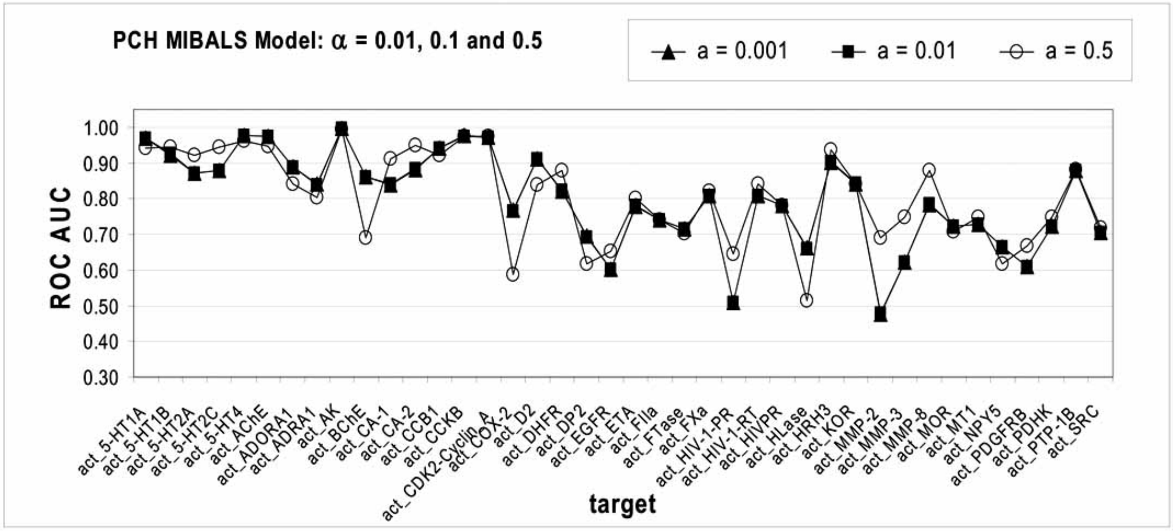


Fig. (13). PCH MIBALS models: Average ROC AUC v. target (*n=15* reference structures) - effect of α (α = 0.001, 0.01, 0.5).

tions. The crucial catechol group oxygen atoms are assigned the highest score by both the PCH and MACCS fingerprints.

MIBALS Congeneric Series Analysis

The PCH and MACCS *core* model fragment scores, as calculated with the above MIBALS model of PDE4 activity, are shown in Fig. (**14**) for the two sample PDE4 inhibitors.

Since this is a core model, the {**F**} set was empty during the model training. Thus, the fragment scores simply reflect the molecular core of the series. Note that similar scores are produced for both structures **1** and **2**, since no activity difference was applied during model training. Almost all molecules in this series contain the catechol group, as reflected by the large PCH and MACCS fragment scores. The carbonyl group on a nitrogen heterocycle is also a common feature in this series, and this is also reflected in the PCH and MACCS fragment scores. In contrast, fused cyclohexenyl ring is not common to the entire series, and is thus assigned relatively low scores in the example compounds. Furthermore, the pendant cyclohexyl group in compound **1**, also not a core feature of this series, is also assigned relatively low scores by both PCH and MACCS fingerprints.

The PCH and MACCS *select* model fragment scores are shown in Fig. (**15**) for the two sample PDE4 inhibitors.

Since this is a select model, the {**F**} set consisted of all training set compounds with $pIC_{50} < 7$. Thus, the fragment scores reflect differences between the congeners that either enhance activity (positive score, solid sphere) or diminish activity (negative score, wire sphere) with respect to the core structure. Comparison of the select fragment scores with the core fragment scores in the previous figure reveals a general

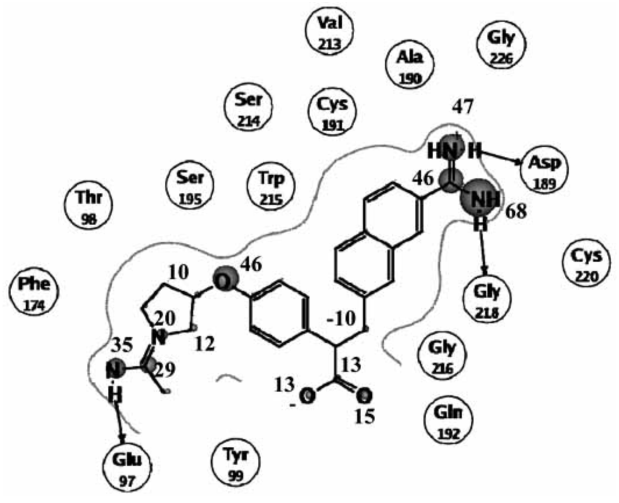


Fig. (14). MACCS fingerprint MIBALS FXa pharmacophore.

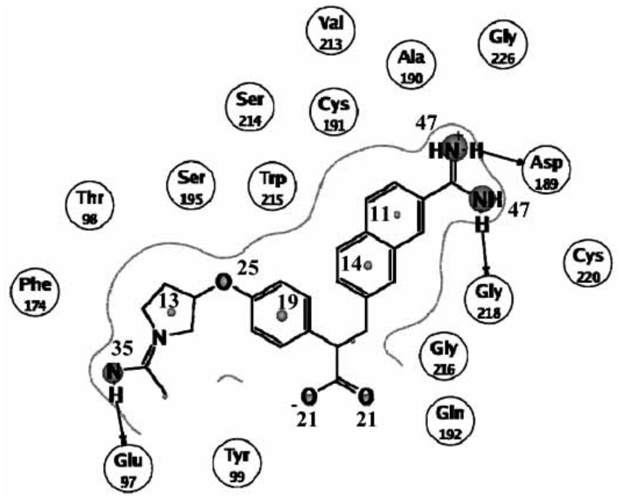


Fig. (15). PCH Fingerprint MIBALS FXa pharmacophore.

diminishment of the core fragment scores, and an increase in the magnitude of peripheral fragment scores. Compared with the core model fragment scores, the select PCH fragment scores for compound **1** show decreases in the core feature scores (e.g. the catechol and the carbonyl) and an increase in the pendant cyclohexyl score. The PCH select fragment scores for compound **2** show diminished core fragment scores (compared with the core model fragment scores), coupled with the appearance of a large negative sphere on the unsubstituted heterocycle NH group. Indeed, many compounds in this series which contain an unsubstituted NH group exhibit diminished activity. The MACCS select fragment scores for compounds **1** and **2** are markedly different from the MACCS core model fragment scores. In both case the MACCS scores for the core molecular fragments are near zero, reflecting the presence of these features in both the active and inactive subsets. The MACCS select scores correctly identify the fluoro groups and the unsubstituted NH group in compound **2** as potential problem points. The MACCS select scores for compound **1** further highlight the importance of substitution on the heterocyclic trivalent N. The unsubstituted NH group gets a negative score in compound **2**. However, in compound **1**, this nitrogen is substituted, and is assigned a large positive score.

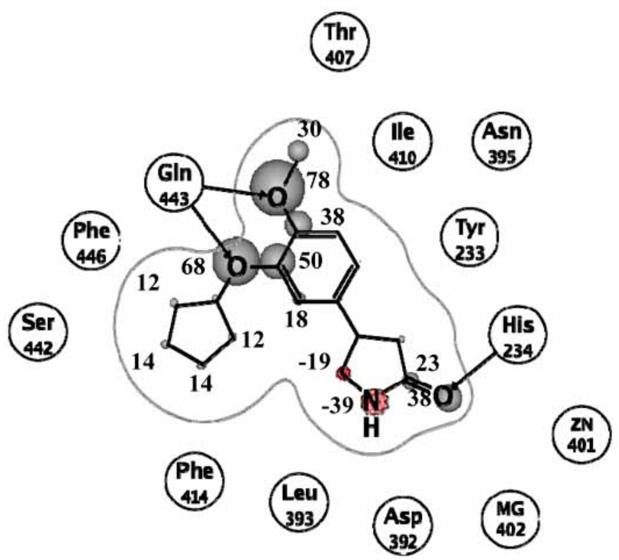


Fig. (16). MACCS MIBALS pharmacophore model for PDE4.

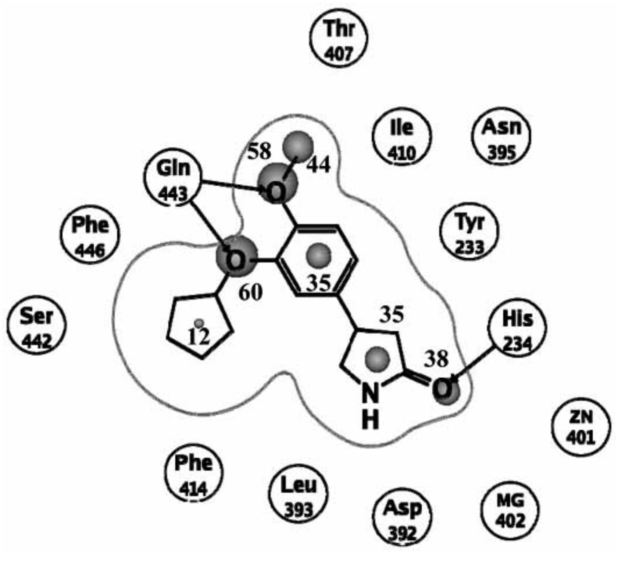


Fig. (17). PCH MIBALS pharmacophore model for PDE4 inhibitors.

CONCLUSIONS

In this work we have developed a method for scoring molecules based on mutual information analysis of 2D fingerprints. The method was validated by application to vHTS experiments, pharmacophore elucidation and congeneric series analysis. Although the validation studies presented here are not exhaustive, these preliminary results are encouraging because they indicate that the method is capable of

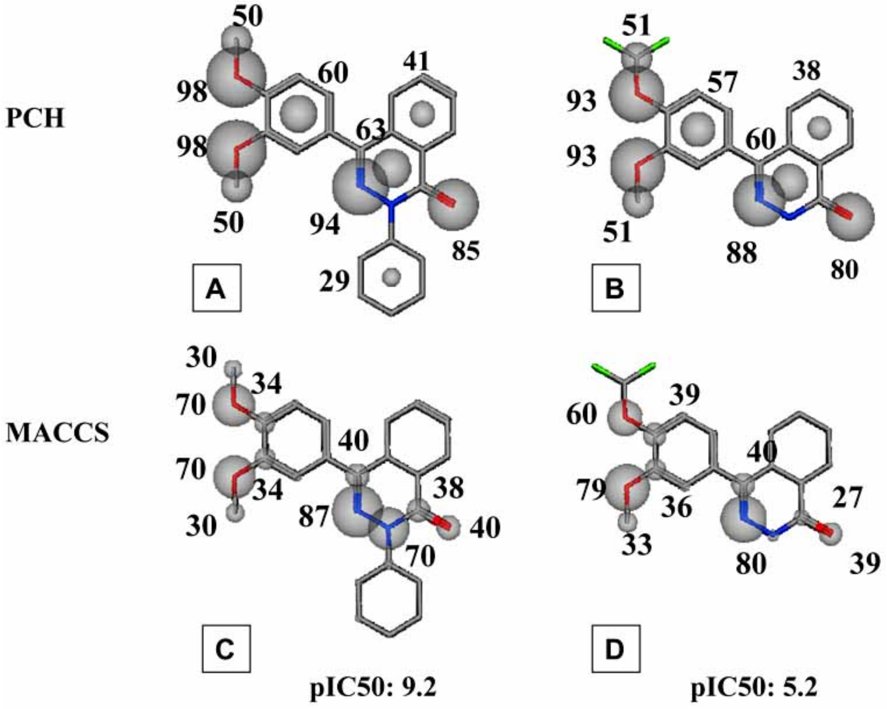
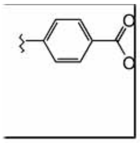


Fig. (18). *Core* PDE4 MIBALS models for compounds **1** and **2;** Fragment scores for 'active' compound **1** (pIC$_{50}$ = 9.2) using the core PCH model (**A**) and the core MACCS model (**C**). Fragment scores for 'inactive' compound **2** (pIC$_{50}$ = 5.2) using the core PCH model (**B**) and the core MACCS model (**D**).



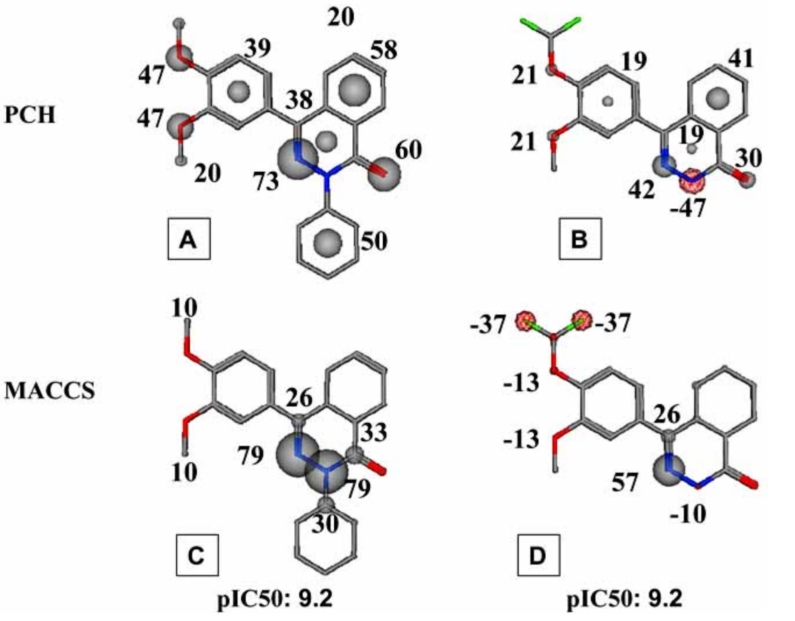


Fig. (19). *Select* PDE4 MIBALS models for compounds **1** and **2**; Fragment scores for 'active' compound **1** (pIC$_{50}$ = 9.2) using the core PCH model (**A**) and the core MACCS model (**C**). Fragment scores for 'inactive' compound **2** (pIC$_{50}$ = 5.2) using the core PCH model (**B**) and the core MACCS model (**D**).

producing reasonable whole molecule scores for virtual screening, as well as meaningful atom scores for important pharmacophore fragments. In addition, this approach can be applied to any fingerprint system where it is possible to map between the fingerprint bits and the atoms used in their discussion.

ACKNOWLEDGEMENTS

We would like to acknowledge J. Christian Baber, Todd Ewing, Yinghong Gao, Miklos Feher, and Gerald Maggiora for helpful discussions and general encouragement.

REFERENCES

[1] Brown, R.; Martin, Y.C. *J. Chem. Inf. Comput. Sci.*, **1996**, *36*, 572-584.

[2] Durant, J.L.; Leland, B.A.; Henry, D.R.; Nourse, J.G. *J. Chem. Inf. Comput. Sci.*, **2002**, *42*, 1273-1280.

[3] ECFP*/FCFP*, Extended Connectivity Rings, Scitegic Inc., San Diego CA: USA 92123 www.scitegic.com

[4] Bender, A.; Glen, R.C. *J. Chem. Inf. Comp. Sci.*, **2002**, *44*, 1708-1718.

[5] Xue, L.; Godden, J.W.; Bajorath, J. *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 881-886.

[6] BCI - Barnard Chemical Information Ltd., Sheffield: UK, www.bci.gb.com

[7] Sheridan, R. P.; Miller, M.D.; Underwood, D.J.; Kearsley, S.K. *J. Chem. Inf. Comput. Sci.*, **1996**, *36*, 128-135.

[8] Clark, R.D.; Fox, P.C.; Abrahamian, E.J. In *Virtual Screening in Drug Discovery*; Alvarez, J.; Shoichet, B.; Eds. Taylor and Francis, New York, **2005**.

[9] Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. *Angew. Chem. Int. Ed.*, **1999**, *38*, 2894-2896.

[10] Willett, P. *J. Med. Chem.*, **2005**, *38*, 4183-4199.

[11] Stahl, M.; Mauser, H. *J. Chem. Inf. Model.*, **2005**, *45*, 542-548.

[12] Pozzan, A. *Curr. Pharm. Des.*, **2006**, *12*, 2099-2110.

[13] Shemetulskis, N.E.; Weininger, D.; Blankley, C.J.; Yang, J.J.; Humblet, C. *J. Chem. Inf. Comput. Sci.*, **1996**, *36*, 862-871.

[14] Xue, L.; Godden, J.W.; Stahura, F.L.; Bajorath, J. *J. Chem. Inf. Comp. Sci.*, **2003**, *43*, 1151-1157.

[15] Schneider, G.; Byvatove, E. *J. Chem. Inf. Comp. Sci.*, **2004**, *44*, 993-999.

[16] Thomsen, M.; Dobel, S.; Lassen, P.; Carlsen, L.; Mogensen, B.B.; Hansen, P.E. *Chemosphere*, **2002**, *49*, 1317-1325.

[17] Daylight Chemical Information Systems, Inc. Mission Viejo: CA USA, www.daylight.com

[18] Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley and Sons: New York, **1991**.

[19] Kullbeck, S.; Leibler, R.A. *Ann. Math. Stat.*, **1951**, *22*, 79-86.

[20] Shannon, C.E.; Weaver, W. *The Mathematical Theory of Communication;* University of Illinois Press: Urbana, Il, **1949**.

[21] Venkatraman, V.; Dalby, A.R.; Yang, Z.R. *J. Chem. Inf. Comput. Sci.*, **2004**, *44*, 1686-1692.

[22] Bonchev, D. *Information Theoretic Indices for Characterization of Chemical Structures;* Research Studies Press: Chichester: UK **1983**.

[23] Maggiora, G. M.; Shanmugasundaram, V. *J. Math. Chem.*, **2005**, *38*, 1-20.

[24] Rossi, F.; Lendasse, A.; Francois, D.; Wertz, V.; Verlaysen, M.; *Chemom. Intell. Lab Syst.*, **2006**, *80*, 215-226.

[25] Yockey, H.P. *Information Theory and Molecular Biology*, Cambridge University Press: Cambridge, **1992**.

[26] Sun, H. *J. Med. Chem.*, **2005**, *48*, 4031-4039.

[27] Sun, H. *Chem. Med. Chem.*, **2006**, *1*, 315-322.

[28] Klon, A.E.; Glick, M.; Davies, J.W. *J. Chem. Inf. Comput. Sci.*, **2004**, *44*, 2216-2224.

[29] Ginn, C.M.R.; Willett, P.; Bradshaw, J. *Perspect. Drug Discov. Des.*, **2000**, *20*, 1-16.

[30] Hert, J.; Willett, P.; Wilton, D.J.; Acklin, P.A.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. *Org. Biomol. Chem.*, **2004**, *2*, 3256-3266.

[31] Ginn, C.M.R. *The Application of Data Fusion Methods to Similarity Searching of Chemical Databases;* Ph. D. thesis: University of Sheffield: Sheffield, **1998**.

[32] Jubilant Biosys Ltd., #96, Second Stage Industrial Area, Yeshwanthpur, Bangalore 560 022 India.

[33] MOE (The Molecular Operating Environment) Version 2006.08, software available from Chemical Computing Group Inc., 1010 Sherbrooke Street West, Suite 910, Montreal: Canada H3A 2R7, http://www.chemcomp.com

[34] MDL Information Systems, Inc. www.mdli.com

[35] Williams, C. *Mol. Div*., **2006**, *10*, 311-332.

[36] Henderson, A.R. *Ann. Clin. Biochem.*, **1993**, *30*, 521-539.

[37] Hanley, J.A.; McNeil, B.J. *Radiology*, **1982**, *143*, 29-36.

Received: April 30, 2008 **Revised: July 7, 2008** **Accepted: August 29, 2008**