# METAPRINT: A Metabolic Fingerprint. Application to Cassette Design for High-Throughput ADME Screening

György M. Keserű* and László Molnár

Computer Assisted Drug Discovery, Gedeon Richter Ltd., H-1475 Budapest, Hungary

Received October 12, 2001

METAPRINT, a metabolic fingerprint, has been developed by predicting metabolic pathways and corresponding potential metabolites. Calculated drug-likeness parameters (log *P* and MW) have been incorporated into METAPRINT to allow the encoding of metabolic diversity within a chemical library. The application of METAPRINT in the design of cassette dosing experiments is demonstrated using a library of α-1a antagonists synthesized at Glaxo Wellcome. Results obtained by Ward's clustering algorithm suggest that METAPRINTs are able to discriminate between low- and high-clearance compounds. Cassette design was performed by maximizing the intracassette Euclidean distances between compounds in METAPRINT space, using simulated annealing. Calculated distances in METAPRINT space were in accordance with experimental data.

## INTRODUCTION

Early phase ADME (absorption, distribution, metabolism, and excretion) studies have become an important part of modern drug discovery.[1,2] In vitro and in vivo techniques are applied to investigate previously synthesized compounds, while in silico approaches are utilized for virtual screening. In vitro metabolic studies involve the determination of metabolic stability, screens for inhibitors of specific cytochrome P450 isozymes, and identification of the most important metabolites. In vivo measurements are used to predict hepatic metabolic clearance, volume of distribution, and bioavailability. Major metabolites are usually identified and quantified at this stage as well. Although a number of methods are available for the calculation of absorption and distribution profiles,[3] the number of truly predictive metabolic models is limited. Present approaches could be divided into three main classes: (i) QSAR and pharmacophore models;[4,5] (ii) protein models;[6−8] (iii) expert systems.[9,10] QSAR and pharmacophore models are developed to predict substrates and inhibitors of a specific cytochrome P450 isozyme. Protein models are employed in low- or high-throughput docking studies and are used to rationalize metabolite formations and identify possible substrates or inhibitors. A detailed analysis of interactions formed within the active site sometimes allows the prediction of potential metabolites as well. Stereoelectronic factors involved in metabolic transformations can be taken into account using quantum chemical calculations. Expert systems are predictive databases that attempt to identify potential metabolites of a compound on the basis of knowledge of rules defining the most likely products.

The underlying premise in these approaches is the existence of a structure−metabolism relationship (SMR) that can be explored by (i) ligand-based, (ii) structure-based, or (iii) knowledge-based techniques. These methods, however, suffer from the paradox pointed out by Testa:[11] the greater the chemical diversity of the investigated compounds, the smaller the chance that SMRs exist and can be uncovered. On the other hand, the information content of an SMR—if it exists—will increase as the boundaries of the chemical space and the diversity of the compounds under investigation increases.

Lead discovery typically starts from a large, structurally diverse set of compounds. Time-consuming and labor-intensive in vivo studies are omitted at this stage, but some pharmacokinetics (PK) groups have argued that in vitro ADME screens are also too expensive to be used in a routine fashion.[12] Although ADME data represent added value for a company offering compound collections, the price for obtaining this information is prohibitively high. Thus, if ADME is to be taken into consideration, in silico approaches represent the only true alternative. The paradox above, however, suggests that development of an SMR would prove difficult at this stage. On the other hand, the application of a SMR developed for a particular series of compounds would not be suitable for virtual screening of diverse chemical libraries.

During the critical hit-to-lead transition we are looking for viable leads having acceptable activity, selectivity, and metabolic profile, both in vitro and in vivo. The chemistry space around the lead compounds is explored during lead optimization. At this stage, relatively small, druglike libraries of structurally similar compounds are generated and evaluated using both in vitro and in vivo methods. The problem reverses here: now it is the information content of the SMR that is limited by the structural similarity of the compounds under investigation. During PK optimization, one should discriminate between compounds with acceptable and unacceptable metabolic profiles. Since the compounds are similar, the information content of an SMR is limited, which in turn results in more limited discriminating power. Effective exploration of the chemistry space around a particular set of leads requires focused libraries considering activity and/or selectivity issues. Efficient PK optimization, however,

* Corresponding author phone: +36-14314605; fax: +36-14326002; e-mail: gy.keseru@richter.hu.

10.1021/ci010106t CCC: $22.00 
Published on Web 02/19/2002

requires metabolic diversity within the focused library that could not be achieved by the application of a simple SMR with limited information content. Although structural keys and 2D fingerprints coupled with hierarchical clustering methods have been effective in separating biologically active from inactive molecules in molecular diversity studies,[13,14] the high degree of structural similarity—especially in combinatorial libraries with a common core—prevents their application in metabolic diversity analyses. In this work, we introduce for the first time a metabolic fingerprint, METAPRINT, for the assessment of metabolic similarity and diversity in combinatorial chemical libraries.

## METHODS

**Computation of METAPRINTs.** Our metabolic fingerprint contains information on predicted routes involved in the metabolism of a given compound and also physicochemical parameters calculated for both the parent compound and its metabolites. Metabolic routes and metabolites of compounds were predicted by the expert system MetabolExpert.[15] MetabolExpert is a rule-based system with 179 rules collected from transformations occurring in mammals. The biotransformation graph created using transformation rules extracted from experimental metabolic trees models the composition of the living system together with transport processes and metabolic pathways. Transformations are formulated as if/then rules that comprise four elements: breaking bonds (active substructure); bonds and atoms introduced (replacing substructures); list of substructures at least one of which must be present (positive condition for the transformation); list of substructures whose presence prevents the transformation (negative condition). The output of MetabolExpert is an estimation of the structural formula of possible metabolites and the metabolic routes involved in their formation. Thus, the first step in the computation of a METAPRINT is extracting the structures of the original compound and the predicted metabolites from MetabolExpert's output file (MDL RD format), followed by calculation of the *c*log*P*[16] and MW for all extracted compounds (using SYBYL 6.6[17] in command line mode driven directly from our Tcl script). After all the data is in hand, the METAPRINT is constructed using the following procedure. The first 179 bins of the METAPRINT contain information on the metabolic transformations involved. Each bin is associated with a particular metabolic reaction and represents the percentage of occurrence of that reaction type in the metabolic tree of the compound of interest. The remaining 201 bins contain metabolite and parent compound *c*log*P* and MW information. The *c*log*P* interval [−10, 10] is divided into 201 bins of equal width (0.1 *c*log*P* units each), and each bin is assigned a value that represents the percentage of the sum of the MW of the metabolites that fall within that particular *c*log*P* range. This complex structure of METAPRINT ensures that relevant drug-likeness and metabolic information is encoded in the fingerprint. These METAPRINTs are able to encode metabolic diversity and can be used in a wealth of applications ranging from simple profiling applications to similarity searching and QSAR. In this paper, we outline the potential use of METAPRINTS in identifying clusters (cassettes) of compounds, which are expected to exhibit minimal overlap in HPLC/MS analysis of samples obtained from in vitro or in vivo metabolic studies of mixtures. The output of the METAPRINT generator program is a simple ASCII file containing one molecule/line, with the name of the molecule and its associated METAPRINT.

**Cassette Design.** In vivo pharmacokinetic parameters are usually determined using mixtures of compounds known as cassettes. Typically, 5−10 compounds are mixed together, and the mixture is administered to animals. After administration, blood samples are collected at prescribed time intervals, and the major metabolites and metabolic stability of the individual compounds are determined. If the co-administered compounds have similar metabolites, then it is very hard to detect them by HPLC/MS techniques. This problem is compounded when the parent compounds also have similar molecular weight. It is therefore important to determine which compounds should be administered simultaneously so that ambiguity is kept to a minimum.

In more formal terms, given a particular cassette size and a collection of compounds considered for PK dosing, the problem is to partition the compounds into the cassettes in a way that maximizes the ability to detect the metabolites by mass-spectrometric techniques. This can be accomplished indirectly by encoding the metabolic profiles of the compounds using METAPRINT fingerprints and maximizing the metabolic diversity of the resulting cassettes. Due to the impact of combinatorial chemistry in modern drug discovery, diversity has been the subject of many studies in the past few years, and several methods have been devised to encode this concept, both in a chemical and mathematical sense (for recent reviews, see refs 18 and 19).

In this work, the metabolic diversity of a given set of compounds is measured by the minimum pairwise Euclidean distance of their respective METAPRINTs.[20] The objective is to partition the compounds among the cassettes in a way that the compounds within each cassette are as different from each other as possible. This is a difficult combinatorial problem, since the number of possible solutions (partitions of $n$ compounds into subsets of size $k$) is

$$\prod_{j=0}^{<n/k} \frac{(n-(jk))!}{k!(n-k(j+1))!} \quad n > k; \quad n \bmod k = 0$$

If the number of compounds cannot be evenly divided by the cassette size, the last cassette will contain fewer compounds. For $n = 20$ and $k = 5$, the number of possible combinations is $\sim 1.7 \times 10^{10}$. Clearly, to find the optimal or a nearly optimal solution, a global search algorithm capable of overcoming local minima is required. The metropolis Monte Carlo simulated annealing (MMCSA) method was chosen here because of its programmatic simplicity and ability to handle virtually any conceivable objective function. This method has been used successfully in the past to maximize molecular diversity[21,22] and design combinatorial libraries using complex multiobjective fitness functions.[23−25]

The method starts from an initial random state and walks through the multidimensional space associated with the objective function by small random steps. In this particular problem, a state represents a particular partitioning of $k$ molecules into $n/k$ cassettes, and a step is a small random change in the partitioning by means of an exchange of compounds between two randomly chosen cassettes. In our

implementation, the size of a step (i.e. the number of molecules to be exchanged between two cassettes) was proportional to the simulated annealing pseudotemperature. This approach is somewhat unconventional (in "classical" annealing, the step is local in nature) and was designed to allow the algorithm to cover a wider area of the state space in a finite number of steps at higher temperatures. In MMCSA minimization, a downhill movement is always accepted, while uphill movements are accepted with a probability that depends on the difference in the fitness between the new and the old states. This probability is computed according to the following function:

$$p = e^{-\Delta f/cT}$$

Here $\Delta f$ is the difference of the objective function values between the current and the previous state, $T$ is the pseudotemperature, and $c$ is a constant that is used to scale the acceptance ratio (in the original Metropolis' acceptance criterion, $c$ is equal to Boltzmann's constant and $\Delta f$ is equal to $\Delta E$).

Since the optimal design must place the most metabolically dissimilar compounds in each cassette, we used the objective function

$$-\min_{s}\left(\min_{i,j}^{j=i+1}(d_{ij})\right)$$

where $d_{ij}$ is the Euclidean distance in METAPRINT space between compounds $i$ and $j$ in cassette $s$. This function is similar to the maximum pairwise dissimilarity used in combinatorial library design.

**Computational Details.** All calculations were carried out on a Silicon Graphics Origin 2000 server running IRIX 6.5.11, using programs written by the authors in standard Tcl/Tk. In our algorithm, the triangular distance matrix (which contains only the matrix elements above the diagonal) was computed first for all pairs of compounds to eliminate redundant computation in evaluating the objective function. Once the distance matrix was computed, the optimization was carried out in approximately 10 min on a MIPS 10000 processor. The reader should be aware that these timings reflect the interpreted nature of the Tcl language and that the computational requirements of the algorithm could be dramatically reduced if the programs were implemented in a compiled language such as C or C++.

We used a modified temperature program (Figure 1).[26] This cooling/heating schedule is similar to the tempering method used to harden Damascus blades. It ensures that the minimization will not be frozen in a high-energy local minimum (an amorphous, energetically metastable state) which can occur in a simple cooling schedule. Cassette optimization was performed in 5 annealing cycles (Nann), using 500 MMC steps at each temperature cycle, an initial temperature of 10 000, and a temperature reduction factor of 0.9 (to a final temperature of 10).

## RESULTS AND DISCUSSION

**Data Set.** The data set used in this study is a library of 65 structurally related α-1a antagonists (Chart 1) synthesized at Glaxo Wellcome.[27] These compounds were originally selected for the first cassette dosing PK experiment reported in 1998.[28] Although it was never stated either in patents[29] or in subsequent publications[30−32] that these compounds have been prepared by combinatorial chemistry, the compounds share a common core with variations introduced at three different sites. The structural diversity of this library was assessed using Tanimoto coefficients based on 2D Unity fingerprints. The maximum, mean, and minimum Tanimoto coefficients were found to be 1.00, 0.79, and 0.63, respectively, indicating a high degree of structural similarity which is typical of combinatorial libraries. Possible metabolic routes and metabolites of all individual compounds were first predicted by MetabolExpert. *c*log*P* [33] and molecular weight values were calculated for each metabolite and were compressed, along with the metabolic pathways, into a 380-dimensional METAPRINT vector for each antagonist.

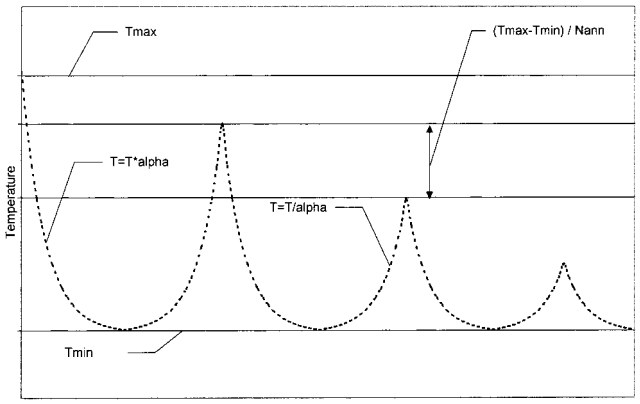
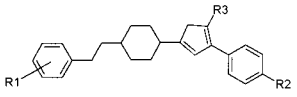



**Figure 1.** Temperature cooling/heating cycles used in MMCSA.

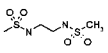
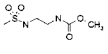
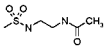
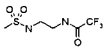
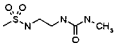
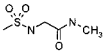
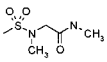
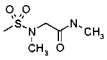
**Evaluation of Metabolic Diversity.** All of the 65 compounds in the test library have been evaluated on an in vivo PK screen at Glaxo Wellcome. Data obtained after simultaneous administration of 5−22 compounds of this congeneric series to dogs were used to derive structure−PK relationships. Clearance, elimination half-life, and volume of distribution were measured for each compound, and the results confirmed this library to be pharmacokinetically diverse. The distributions of PK parameters are depicted in Figure 2. These distributions suggest that it should be possible to discriminate these compounds if appropriate descriptors were used. To demonstrate the discriminating power of our metabolic fingerprint, we applied multidimensional scaling[34] to project the 380-dimensional METAPRINT space to a 2-dimensional plane and compare the corresponding map to that obtained from the 992-dimensional 2D Unity fingerprints (Figure 3). This method has been very effective in visualizing diversity space using a wide spectrum of molecular descriptors.[35] The stress values of the 2-dimensional projections were found to be 0.281 and 0.088, respectively. This difference may be attributed to the binary nature, and therefore reduced information content, of the Unity fingerprints. These maps confirm that the compounds are well separated in METAPRINT space, which is the intent of the algorithm. Moreover, assuming that the Unity fingerprints are a sufficiently good measure of structural diversity, these maps, along with the distributions of the experimental data, also suggest that conventional structural descriptors may not be suitable for encoding metabolic diversity.

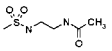
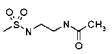
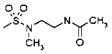
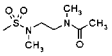
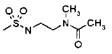
**Cluster Analysis.** To analyze the degree of separation between compounds with low and high clearance, we

**Chart 1**




General formula

| Molecule name | R1 | R2 | R3 |
|---|---|---|---|
| 1 | 4H | OMe | H |
| 2 | 4H | OMe |  |
| 3 | 4H | OMe |  |
| 4 | 4H | OMe |  |
| 5 | 4H | OMe |  |
| 6 | 4H | OMe |  |
| 7 | 4H | OMe |  |
| 8 | 4H | OMe |  |
| 10 | 4H | OMe |  |
| 11 | 4H | OMe |  |
| 12 | 4F | OMe |  |
| 13 | 4H | OMe |  |
| 14 | 4H | OMe |  |
| 15 | 4F | OMe |  |
| 16 | 4F | OMe | H |
| 17 | 4F | OMe | F |

| Molecule name | R1 | R2 | R3 |
|---|---|---|---|
| 35 | 4H | OEt |  |
| 36 | 4F | OEt |  |
| 37 | 4F | OEt |  |
| 38 | 4F | OEt |  |
| 39 | 4F | OEt |  |
| 40 | 4F | H | H |
| 41 | 4F | F | H |
| 42 | 4F | F | F |
| 43 | 4F |  | H |
| 44 | 4F |  | H |
| 45 | 4F |  | H |
| 46 | 4F |  | H |
| 47 | 4F |  | H |
| 48 | 4F | $NH_2$ | H |
| 49 | 4F | NMe | H |
| 50 | 4F |  | H |

| Molecule name | R1 | R2 | R3 |
|---|---|---|---|
| 18 | 4H | OMe |  |
| 19 | 4F | OMe |  |
| 20 | 3F | OMe |  |
| 21 | 4Cl | OMe |  |
| 22 | 3F,4Me | OMe |  |
| 23 | 2F,3F | OMe |  |
| 24 | 2F,5F | OMe |  |
| 25 | 2F,4F | OMe | H |
| 26 | 3F,4F | OMe |  |
| 27 | 2F,4F | OMe |  |
| 28 | 2F,3F,4F | OMe |  |
| 29 | 2F,4F,5F | OMe |  |
| 30 | 4H | OEt |  |
| 31 | 4F | OEt |  |
| 32 | 2F,4F | OEt |  |
| 33 | 4F | OEt |  |
| 34 | 4F | OEt |  |

| Molecule name | R1 | R2 | R3 |
|---|---|---|---|
| 51 | 4F | $NO_2$ | H |
| 52 | 4F |  | H |
| 53 | 4F |  | H |
| 54 | 4F |  | H |
| 55 | H |  | H |
| 56 | 4F |  | H |
| 57 | 4F |  | H |
| 58 | 4F |  | H |
| 59 | 4F |  | H |
| 60 | 4F |  | H |
| 61 | 4F |  | H |
| 62 | 4F |  | H |
| 63 | 4F |  | H |
| 64 | 4F |  | H |
| 65 | 4F |  | H |
| 66 | 4F |  | H |

clustered the library using the METAPRINT descriptors and Ward's hierarchical clustering algorithm. With a clearance threshold of 5 mL/(min/kg), there are 16 and 50 compounds with low and high clearances, respectively. Following the method published by Martin et al.,[13] we defined low-clearance clusters (LCC) as ones in which at least one member of the cluster has low clearance (LC). The subset of structures which appear in LCCs is denoted as the LCC subset. The difference in the proportion of LC compounds in the LCC subset to that of LC compounds in the whole library is an indication of the degree of separation between low- and high-clearance compounds in the given set of clusters. LC singletons were excluded from this analysis, since their proportion is 1.0 and could skew the results. This analysis has been performed using both experimental and random clearance data to demonstrate that the observed discrimination is the result of clustering rather than the particular distribution of clearance data.

A comparison of random points to the baseline indicates the enrichment of LC compounds; however, using experimental clearance data the increase in the proportion of LC compounds is significantly higher (Figure 4). This further separation is the result of grouping LC compounds together, and therefore, their similarity to other LC compounds is higher compared to high-clearance compounds. This is a clear indication that METAPRINT does contain relevant metabolic information.

**Cassette Design Experiments.** Simultaneous quantitative in vitro and in vivo PK analysis of drugs has been used extensively in early phase drug discovery. Cassette dosing was invented at Glaxo Welcome in the early 1990s,[28] and all of the major pharmaceutical companies at least tested this methodology to increase the speed of their preclinical PK efforts.[36−39] Successful cassette dosing experiments require careful design of the cassettes to answer at least the following questions:

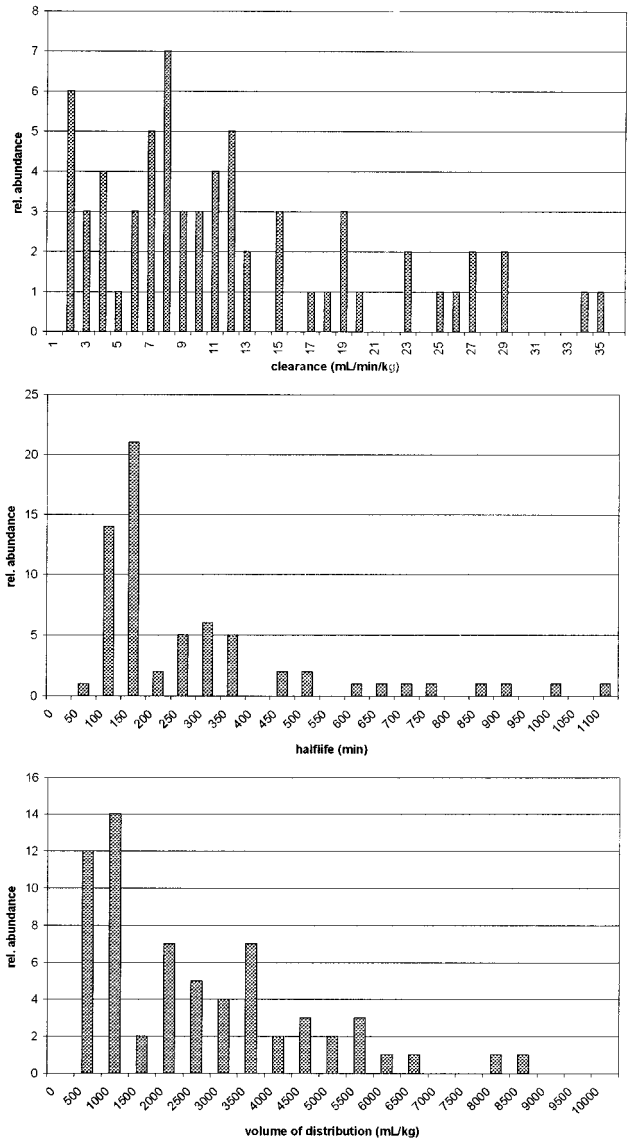



**Figure 2.** Distributions of PK parameters.

(i) How many compounds should be put in each cassette?

(ii) Which compounds should be grouped together?

(iii) What is the maximal dose of compounds that could be administered without adverse effects?

(iv) How sensitive is the assay?

Although answers to iii and iv require domain-specific PK knowledge, questions i and ii have a significant cheminformatic component. Since samples from a cassette dosing experiment are usually analyzed by HPLC/MS, it is generally accepted that isobaric compounds as well as isobaric metabolites should be avoided.[30] Multiple reaction monitoring (MRM)[40] adds great selectivity in monitoring multicomponent samples; however, this selectivity can be significantly reduced when samples are drawn from a mixture of closely related compounds prepared by combinatorial chemistry. These compounds and their metabolites are likely to have common structural features and, consequently, similar fragmentation patterns. To avoid these problems, Frick et al.[41] developed a simple algorithm for calculating the cassette (CTC). Starting from the molecular formula of parent compounds, CTC calculates molecular ions and uses an undisclosed iterative procedure to minimize the number of

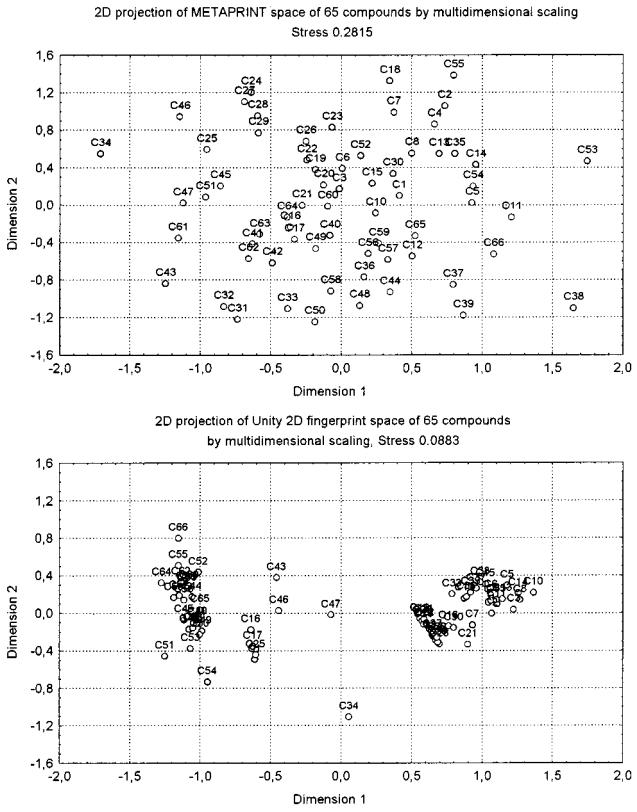



**Figure 3.** 2D projections of METAPRINT and Unity 2D fingerprint spaces.

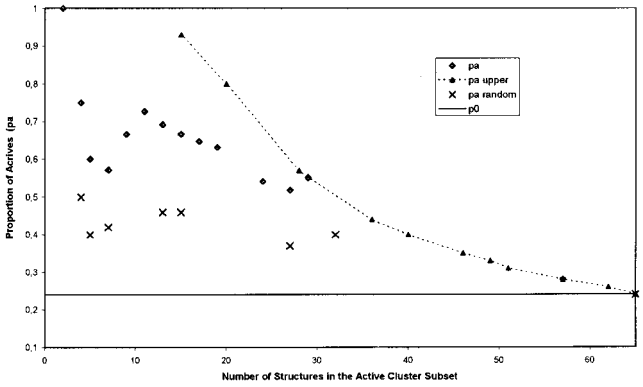



**Figure 4.** Proportion of actives in active cluster subset.

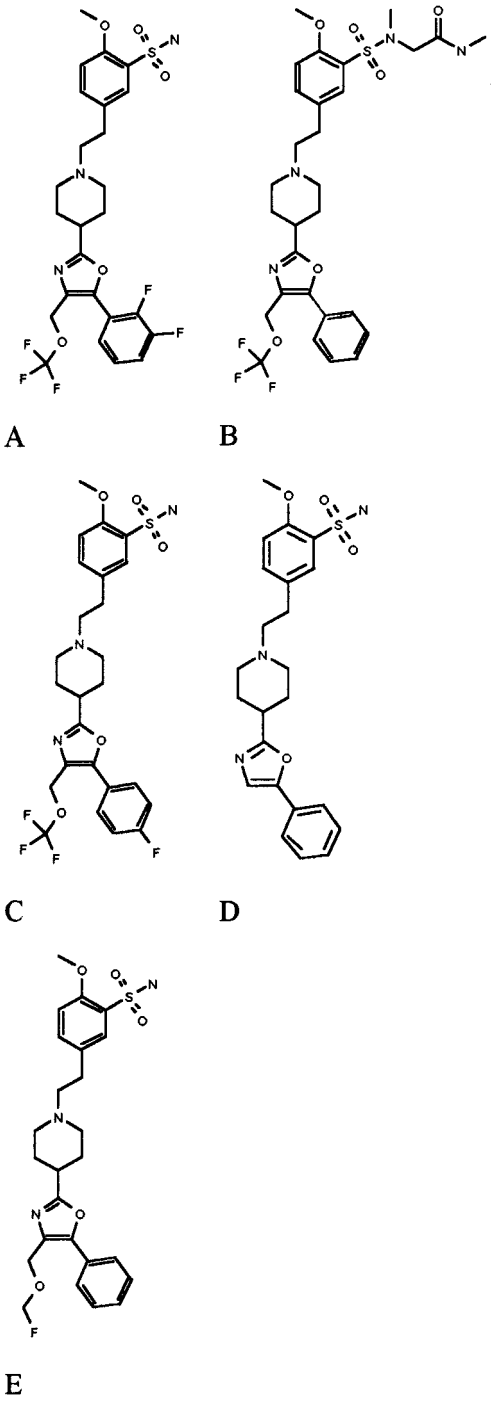
isobaric compounds within each cassette. They also excluded compounds that differ by critical molecular weights such as 1 (arising from rare isotopes of carbon and hydrogen) and 16 (equivalent to the addition of oxygen). One of the major drawbacks of this calculation is that potential metabolites were not included in the design process. Since METAPRINTs contain information for both the parent compounds and their putative metabolites that are relevant to HPLC/MS, we tested our metabolic fingerprint in cassette design. Since we could not obtain quantitative information on metabolic profiles, we considered the worst case where all of the potential metabolites of all compounds in the study are formed in equal amounts. According to the recent cassette dosing literature, we have focused on five-membered cassettes since the throughput cannot be significantly increased using smaller cassette sizes ($n < 5$). On the other hand, as the cassette size increases ($n > 10$), the probability of drug−drug interactions increases.[41] In general, the optimal cassette

**Table 1.** Cassette Configurations

| cassette name | cassette members | | | | | minimal intracassette dist |
|---|---|---|---|---|---|---|
| Initial Cassette Configuration | | | | | | |
| cassette_no. 1 | 8 | 7 | 66 | 65 | 64 | 44.01 |
| cassette_no. 2 | 63 | 62 | 61 | 60 | 6 | 39.45 |
| cassette_no. 3 | 59 | 58 | 57 | 56 | 55 | 35.55 |
| cassette_no. 4 | 54 | 53 | 52 | 51 | 50 | 44.20 |
| cassette_no. 5 | 5 | 49 | 48 | 47 | 46 | 41.71 |
| cassette_no. 6 | 45 | 44 | 43 | 42 | 41 | 36.46 |
| cassette_no. 7 | 40 | 4 | 39 | 38 | 37 | 26.75 |
| cassette_no. 8 | 36 | 35 | 34 | 33 | 32 | 31.22 |
| cassette_no. 9 | 31 | 30 | 3 | 29 | 28 | 36.83 |
| cassette_no. 10 | 27 | 26 | 25 | 24 | 23 | 14.63 |
| cassette_no. 11 | 22 | 21 | 20 | 2 | 19 | 16.78 |
| cassette_no. 12 | 18 | 17 | 16 | 15 | 14 | 34.09 |
| cassette_no. 13 | 13 | 12 | 11 | 10 | 1 | 39.69 |
| Final Cassette Configuration | | | | | | |
| cassette_no. 1 | 52 | 42 | 19 | 21 | 8 | 42.22 |
| cassette_no. 2 | 45 | 35 | 46 | 37 | 25 | 46.08 |
| cassette_no. 3 | 27 | 2 | 10 | 4 | 66 | 43.48 |
| cassette_no. 4 | 55 | 43 | 41 | 28 | 36 | 47.39 |
| cassette_no. 5 | 7 | 61 | 6 | 16 | 24 | 43.23 |
| cassette_no. 6 | 51 | 38 | 26 | 65 | 3 | 41.68 |
| cassette_no. 7 | 31 | 60 | 32 | 20 | 62 | 42.97 |
| cassette_no. 8 | 53 | 63 | 23 | 34 | 58 | 47.68 |
| cassette_no. 9 | 56 | 5 | 50 | 29 | 40 | 43.77 |
| cassette_no. 10 | 12 | 39 | 18 | 44 | 15 | 43.50 |
| cassette_no. 11 | 49 | 57 | 1 | 14 | 22 | 42.74 |
| cassette_no. 12 | 17 | 13 | 48 | 33 | 11 | 46.75 |
| cassette_no. 13 | 59 | 64 | 47 | 54 | 30 | 42.70 |

size depends strongly on the chemical characteristics of the compounds studied. Using the protocol described here, one can adjust the optimal cassette size by trial-and-error, searching for the largest cassette size with nonviolating members. A total of 65 compounds from the α-1a antagonist library were grouped randomly to 13 cassettes as a starting configuration (Table 1). Minimal and maximal distances within cassettes in the METAPRINT space were 14.63 and 44.00, respectively. The MMCSA procedure generated and evaluated 150 000 configurations. Simulated annealing performed in 300 temperature steps resulted in a cassette configuration with a minimal and maximal distance of 41.68 and 47.68, respectively. Unfortunately, Glaxo did not disclose the cassette configuration they used; our design process, however, identified a possible cassette configuration for this set of compounds. Validation of our approach could only be done using another study reporting five compounds that were measured simultaneously in vivo (Chart 2).[31] Two-dimensional nonlinear maps of this cassette were generated using both METAPRINT and Unity fingerprints, in the manner described above (Figure 5). Again, in contrast to the Unity fingerprints, the compounds are well-separated in METAPRINT space, which is in accordance with the results of the cassette dosing experiment.

**Chart 2**




## CONCLUSIONS

In this work, we presented a metabolic fingerprint for organic compounds on the basis of the prediction of metabolic pathways, structure, and drug-likeness of their corresponding metabolites. METAPRINTs were utilized for analyzing the metabolic diversity of a library of α-1a antagonists. In contrast to conventional 2D Unity fingerprints, we found that compounds with different PK parameters can be discriminated in METAPRINT space. Ward's clustering of the compounds in the library demonstrated the capability of METAPRINTs to differentiate between low- and high-clearance compounds. These calculations revealed that METAPRINTs contains relevant metabolic information that can be used in the design of metabolically diverse libraries. We demonstrated that similarity and diversity measures based on METAPRINT representations can be very effective in designing good cassette dosing experiments. A simulated annealing scheme has been developed to minimize the overlap of parent compounds and their metabolites during the HPLC/MS analysis of cassette samples. Two-dimensional nonlinear maps of METAPRINT spaces calculated for compounds studied in a single cassette were in accordance

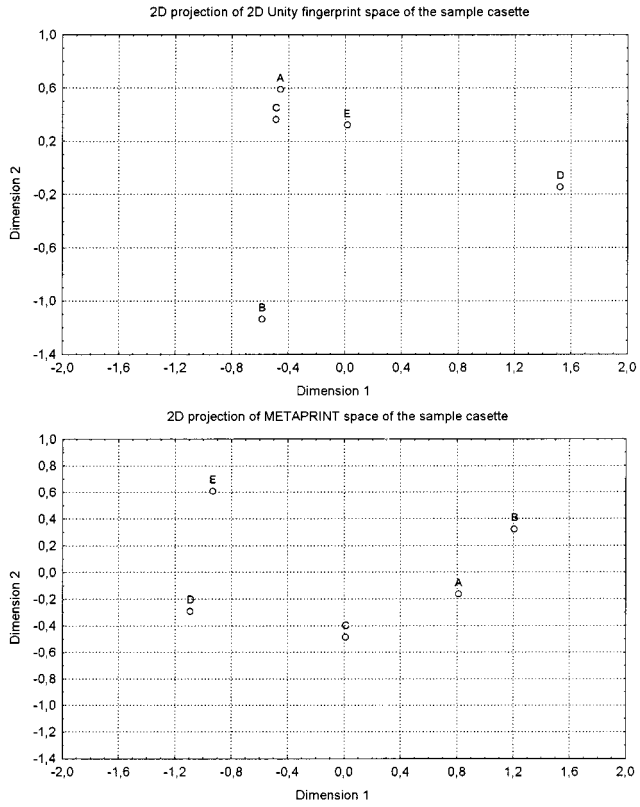



**Figure 5.** 2D projections of METAPRINT and Unity 2D fingerprint spaces of the sample cassette.

with the experimentally observed nonoverlapping character of the cassette members.

## ACKNOWLEDGMENT

We thank Zoltán Kovári (Gedeon Richter) for carefully reading the manuscript and Dr. Dimitris K. Agrafiotis (3-Dimensional Pharmaceuticals, Inc.) for helpful discussions.

## REFERENCES AND NOTES

(1) Roberts, S. A. High-throughput screening approaches for investigating drug metabolism and pharmacokinetics. *Xenobiotica* **2001**, *31*, 557−589.

(2) Watt, A. P.; Morrison, D.; Evans, D. C. Approaches to higher-throughput pharmacokinetics (HTPK) in drug discovery. *Drug Discovery Today* **2001**, *5*, 17−24.

(3) Matter, H.; Baringhaus, K.-H.; Naumann, T.; Klabunde, T.; Pirard, B. Computational Approaches Towards the Rational Design of Drug-like Compound Libraries. *Comb. Chem. High Throughput Screening* **2001**, *4*, 453−475.

(4) Ekins, S.; Bravi, G.; Blinkley, S.; Gillespie, J. S.; Ring, B. J.; Wikel, J. H.; Wrighton, S. A. Three- and four-dimensional quantitative structure activity relationship analyses of cytochrome P-450 3A4 inhibitors. *J. Pharm. Exp. Ther.* **1999**, *290*, 429−438.

(5) Ekins, S.; Bravi, G.; Blinkley, S.; Gillespie, J. S.; Ring, B. J.; Wikel, J. H.; Wrighton, S. A. Three and four dimensional-quantitative structure activity relationship (3D/4D-QSAR) analyses of CYP2D6 inhibitors. *Pharmacogenetics* **1999**, *9*, 477−489.

(6) De Groot, M. J.; Vermeulen, N. P. Modeling the active sites of cytochrome P450s and glutathione S-transferases, two of the most important biotransformation enzymes. *Drug Metab. Rev.* **1997**, *29*, 747−799.

(7) Keserü, G. M.; Kolossváry, I. *Molecular Mechanics and Conformational Analysis in Drug Design*; Blackwell Science: Oxford, U.K., 2000; pp 147−164.

(8) Keserü, G. M. A virtual high throughput screen for high affinity cytochrome $P450_{cam}$ substrates. Implication for in silico prediction of drug metabolism. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 649−657.

(9) Darvas, F.; Marokházi, S.; Kormos, P.; Kulkarni, P.; Kalász, H.; Papp Á. In *Drug Metabolism, Databases and High Throughput Testing During Drug Design and Development*; Erhardt, P. W., Ed.; Blackwell Science: Cambridge, MA, 1999; pp 237−270.

(10) Klopman, G.; Tu, M. In *Drug Metabolism, Databases and High Throughput Testing During Drug Design and Development*; Erhardt, P. W., Ed.; Blackwell Science: Cambridge, MA, 1999; pp 271−276.

(11) Testa, B.; Cruciani, G. In *Pharmacokinetic Optimization in Drug Research: Biological, Physicochemical and Computational Strategies*; Testa, B., van de Waterbeemd, H., Folkers, G., Eds.; Verlag Helvetica Chimica Acta (VHCA); Wiley-VCH: Zürich; Weinheim, Germany, 2001; pp 65−84.

(12) Eddershaw, P. J.; Beresford, A. P.; Bayliss, M. K. ADME/PK as part of a rational approach to drug discovery. *Drug Discovery Today* **2000**, *5*, 409−414.

(13) Brown, R. D.; Martin, Y. C. Use of structure−activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

(14) Patterson, D. E., Cramer, R. D., Ferguson, A. M., Clark, R. D. and Weinberger, L. E. Neighborhood behavior: a useful concept for validation of molecular diversity descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.

(15) *MetabolExpert*, version 11.0 implemented in Pallas 2.1; ComDrug Ltd.: Budapest, Hungary.

(16) Leo, A. J. Calculating log $P_{oct}$ from structures. *Chem. Rev.* **1993**, *93*, 1281−1306.

(17) *Sybyl 6.6*; Tripos Inc.: St. Louis, MO.

(18) Martin, E. J.; Spellmeyer, D. C.; Critchlow, R. E., Jr.; Blaney, J. M. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: Weinheim, Germany, 1997; Vol. 19.

(19) Agrafiotis, D. K., Myslik, J. C., and Salemme, F. R. Advances in diversity profiling and combinatorial series design. *Mol. Divers.* **1999**, *4*, 1−22.

(20) Lajiness, M. S. In *QSAR: Rational Aproaches to the Design of Bioactive Compounds*; Silipo, C., Vittoria, A., Eds.; Elsevier: Amsterdam, 1991; pp 201−204.

(21) Agrafiotis, D. K. Stochastic algorithms for maximizing molecular diversity. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 841−851.

(22) Hassan, M.; Bielawski, J. P.; Hempel, J. C.; Waldman, M. Optimization and visualization of molecular diversity of combinatorial libraries. *Mol. Divers.* **1996**, *2*, 64−74.

(23) Rassokhin, D. N.; Agrafiotis, D. K.. Kolmogorov-Smirnov statistic and its applications in library design. *J. Mol. Graphics Modell.* **2000**, *18*, 370−384.

(24) Agrafiotis, D. K. Multiobjective optimization of combinatorial libraries. *IBM J. Res. Develop.* **2001**, *45*, 545−566.

(25) Agrafiotis, D. K.; Rassokhin, D. N. Design and prioritization of plates for high-throughput screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 798−805.

(26) *PowerSearch*; Tripos Inc.: St. Louis, MO.

(27) Shaffer, J. E.; Adkinson, K. K.; Halm, K.; Hedeen, K.; Berman, J. Use of "N-in-one" dosing to create an in vivo pharmacokinetics database for use in developing structure−pharmacokinetic relationships. *J. Pharm. Sci.* **1999**, *88*, 313−318.

(28) Berman, J.; Halm, K.; Adkison, K.; Shaffer, J. Simultaneous pharmacokinetic screening of a mixture of compounds in the dog using API LC/MS/MS analysis for increased throughput. *J. Med. Chem.* **1997**, *40*, 827−829.

(29) WO 9616049, 1996; *Chem. Abstr.* **1997**, *125*, 142709.

(30) Frick, L. W.; Adkision, K. K.; Wells-Knecht, K. J.; Woolard, P.; Higton, D. M. Casette dosing: rapid in vivo assessment of pharmacokinetics. *Pharm. Sci. Technol. Today* **1998**, *1*, 12−18.

(31) Cai, Z.; Sinhababu, A. K.; Harrelson, S. Simultaneous quantitative cassette analysis of drugs and detection of their metabolites by high performance liquid chromatography/ion trap mass spectometry. *Rapid Commun. Mass Spectrosc.* **2000**, *14*, 1637−1643.

(32) Cai, Z.; Han, C.; Harrelson, S.; Fung. E.; Sinhababu, A. K. High-throughput analysis in drug discovery: application of liquid chromatography/ion-trap mass spectrometry for simultaneous cassette analysis of α-1a antagonists and their metabolites in mouse plasma. *Rapid Commun. Mass Spectrosc.* **2001**, *15*, 546−550.

(33) *clogP*, ver. 4.0 associated database ver. 18; Biobyte Inc.

(34) *STATISTICA for Windows [Computer program manual]*; StatSoft, Inc.: Tulsa, OK, 1996.

(35) Agrafiotis, D. K.; Lobanov, V. S. Multidimensional scaling and visualization of large molecular similarity tables. *J. Comput. Chem.* **2001**, *22*, 488−500.

(36) McLoughlin, D. A.; Olah, T. V.; Gilber, J. D. A direct technique for the simultaneous determination of 10 drug candidates in plasma by liquid chromatography-atmospheric pressure chemical ionization mass spectrometry interfaced to a Prospekt solid-phase. *J. Pharm. Biomed. Anal.* **1997**, *15*, 1893−1901.

(37) Tong, X.; Ita, I. E.; Wang, J.; Pivnichny, J. V. Characterization of a technique for rapid pharmacokinetic studies of multiple coeluting compounds by LC/MS/MS. *J. Pharm. Biomed. Anal.* **1999**, *20*, 773−784.

(38) Rano, T. A.; Cheng, Y.; Huening T. T.; Zhang, F.; Schleif, W. A.; Gabryelski, L.; Olsen, D. B.; Kuo, L. C.; Lin, J. H.; Xu, X.; Olah, T. V.; McLoughlin, D. A.; King, R.; Chapman, K. T.; Tata, J. R. Combinatorial diversification of indinavir: in vivo mixture dosing of an HIV protease inhibitor library. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 1527−1530.

(39) Ward, K. W.; Proksch, J. W.; Levy, M. A.; Smith, B. R. Development of an in vivo preclinical screen model to estimate absorption and bioavailability of xenobiotics. *Drug Metab. Disp.* **2001**, *29*, 82−88.

(40) McLafferty, F. W. *Tandem Mass Spectrometry*; John Wiley: New York, 1983.

(41) Frick, L. W.; Higton, D. M.; Wring, S. A.; Wells-Knecht, K. Casette dosing: rapid estimation of in vivo pharmacokinetics. *Med. Chem. Res.* **1998**, *8*, 472−477.

CI010106T