

Anatomy of Fingerprint Search Calculations on Structurally Diverse Sets of Active Compounds

Jeffrey W. Godden,[†] Florence L. Stahura,[‡] and Jürgen Bajorath^{*,†}

Department of Life Science Informatics, B-IT, Rheinische Friedrich-Wilhelms-Universität, Görresstrasse 13, D-53113 Bonn, Germany, and Institute for Chemical Genomics, 600 Broadway, Suite 580, Seattle, Washington 98123

Received July 4, 2005

Similarity searching using molecular fingerprints is a widely used approach for the identification of novel hits. A fingerprint search involves many pairwise comparisons of bit string representations of known active molecules with those precomputed for database compounds. Bit string overlap, as evaluated by various similarity metrics, is used as a measure of molecular similarity. Results of a number of studies focusing on fingerprints suggest that it is difficult, if not impossible, to develop generally applicable search parameters and strategies, irrespective of the compound classes under investigation. Rather, more or less, each individual search problem requires an adjustment of calculation conditions. Thus, there is a need for diagnostic tools to analyze fingerprint-based similarity searching. We report an analysis of fingerprint search calculations on different sets of structurally diverse active compounds. Calculations on five biological activity classes were carried out with two fingerprints in two compound source databases, and the results were analyzed in histograms. Tanimoto coefficient (Tc) value ranges where active compounds were detected were compared to the distribution of Tc values in the database. The analysis revealed that compound class-specific effects strongly influenced the outcome of these fingerprint calculations. Among the five diverse compound sets studied, very different search results were obtained. The analysis described here can be applied to determine Tc intervals where scaffold hopping occurs. It can also be used to benchmark fingerprint calculations or estimate their probability of success.

INTRODUCTION

Similarity searching using molecular fingerprints is among the most popular approaches for the identification of novel active compounds.^{1,2} In contrast to many compound classification methods, fingerprint searching can be applied when only a single active compound is available as a template for similarity exploration, a situation frequently faced in drug discovery research. A fingerprint search compares a bit string representation of a known active template molecule with those of database compounds, and bit string overlap, as assessed by various similarity metrics, is used as a measure of molecular similarity.¹ For fingerprint searching, the Tanimoto coefficient (Tc) is the most popular similarity metric.¹ In virtual screening, very large numbers of database molecules must be similarity-ranked, thus creating search problems of significant magnitude.²

Many 2D and 3D structural or property descriptors have been encoded in various fingerprint representations.² Despite differences in design and complexity, all fingerprints ultimately transform molecular similarity analysis into a pairwise comparison of bit patterns of template and database compounds. The effectiveness of the approach generally depends on the ability of fingerprint calculations to relate bit string similarity to biological activity, in accord with the similar property principle,³ and distinguish active from inactive

compounds. In virtual screening applications, this often represents a formidable task because the vast majority of database compounds are inactive (and, thus, potential false positives), which creates substantial background problems.² In this context, the sensitivity of fingerprints is a critical issue and so is the effectiveness of similarity measures. Moreover, in light of the similar property principle, a key question is whether molecules that are most similar in fingerprint space have, indeed, similar biological activity; if so, what similarity metric and threshold values would be reliable indicators of true molecular similarity? Such questions have long been focal points of performance analyses of fingerprint searching.

Despite early attempts to generalize activity-related threshold values of the Tanimoto coefficient,^{4,5} it has proven to be very difficult to define cutoff values that reliably indicate biological activity relationships^{2,6} or to associate predefined similarity threshold values with a specific biological activity.⁶ Clearly, similarity searching is not only influenced by characteristic features of the methods used² but also by significant differences in the way various compound classes respond to similarity search tools.⁷ It is, therefore, not surprising that different comparative analyses of fingerprint-based methods and similarity metrics typically produce different results.^{7–10} In addition, there is often only very limited overlap between compound selections based on the application of different methods applied to the same search problem.^{7,10} Accordingly, various attempts have been made to further improve the predictive ability of fingerprint search calculations.

* Author to whom correspondence should be addressed. Tel.: +49-228-2699-306, fax: +49-228-2699-341, e-mail: bajorath@bit.uni-bonn.de.

[†] Rheinische Friedrich-Wilhelms-Universität.

[‡] Institute for Chemical Genomics.

A number of efforts have focused on tuning fingerprint searching toward the recognition of specific compound classes. Provided that multiple active templates are available, centroid⁸ or consensus¹¹ fingerprints can be calculated for different activity classes. Furthermore, fingerprint calculations can also be scaled in an activity class-dependent manner.^{12,13} These techniques essentially merge or modify fingerprints for specific applications. Other attempts have focused on further refinement of similarity scoring schemes. For example, consensus scoring¹⁴ or data fusion techniques^{15,16} can be applied that combine scores obtained from different scoring schemes or similarity metrics. Techniques that operate at the level of scoring can also be applied if only single templates are available, in contrast to fingerprint modifications.

Taken together, there continues to be very little generality in fingerprint-based similarity searching, and consequently, there is a need for diagnostic tools to better understand the performance of these calculations in specific search situations. This is particularly important for virtual screening applications because of the very large number of potential false positives and the scoring noise that is generated. Some graphical methods have been developed for the analysis of fingerprint calculations including cumulative recall curves,¹⁷ which record the retrieval of correctly identified hits, and similarity search profiles.¹⁸ These profiles monitor fingerprint calculations on multiple template compounds over the entire value range of a chosen similarity coefficient and, in parallel, capture the distribution of correctly detected active molecules and false positives. This technique has also been applied to rationalize performance-increasing effects of fingerprint scaling.¹⁹

Herein, we present an analysis designed to reveal the potential of fingerprint searching using single templates. We have deliberately focused on the study of sets of structurally diverse compounds having similar activity, the identification of which is often called lead hopping²⁰ or scaffold hopping.²¹ Results of the compound test sets we have analyzed demonstrate how case-sensitive and, thus, unpredictable the outcome of fingerprint searching can be. We also show how the choice of source databases for benchmarking purposes can bias the results. The histogram analysis of score distributions reported herein can also be used to estimate the potential for success of virtual screens on a case-by-case basis, given selected search tools and compound databases.

MATERIALS AND METHODS

In this study, we have applied two fingerprints that consist of, or contain, structural fragment descriptors. These types of descriptors and fingerprints are widely used in pharmaceutical research,²² are publicly available, and have been shown to perform well in a number of molecular similarity-oriented applications.^{2,22} We have designed and evaluated a number of fingerprint representations containing structural fragment descriptors.²³ One of the fingerprints used in this study consists of a publicly available set of 166 structural keys²⁴ (MACCS). The other, termed MPMFP, consists of 175 bit positions and combines 114 structural keys with 61 binary-transformed and -encoded molecular property descriptors.²⁵ Approximately half of these 61 binary-encoded

descriptors map various physicochemical properties on molecular surface areas' 2D representations of molecules²⁵ and are, thus, best understood as implicit 3D descriptors. The MPMFP fingerprint has been shown to be an effective similarity search tool in a number of benchmark calculations²⁵ and is also publicly available.²⁶ For similarity searching, these two fingerprints were implemented in the Molecular Operating Environment software (MOE).²⁷ Two different source databases were used, the Available Chemicals Directory (ACD),²⁸ containing approximately 200 000 synthetic compounds and reagents, and a database with approximately 900 000 compounds we collected from offerings of various medicinal chemistry vendors, named MCD. MACCS and MPMFP were calculated for all ACD and MCD compounds. Although these databases, at least MCD, may well contain novel active compounds for the biological activity classes studied here, all database molecules were considered inactive (and, thus, potential false positives) in the context of our similarity search calculations.

Test compounds belonging to five biological activity classes were assembled from the literature and patents. These were gonadotropin releasing hormone agonists (GnRH),^{29–33} growth hormone secretagogue agonists (GHS),^{34–36} melanin-concentrating hormone receptor antagonists (MCH),^{37–38} CCR5 chemokine receptor antagonists (CCR5),^{39–40} and JNK protein kinase inhibitors (JNK).^{41–43} Particular care was taken to select activity classes that contained subsets of structurally distinct compounds. As a conformation beyond visual inspection, Tc matrices were calculated with MACCS for each class to evaluate and confirm the presence of structural diversity. The selected activity classes consisted of between 12 and 91 compounds; example structures are shown in Figure 1. As single templates for fingerprint searching, the most potent compound from each class was selected, as also shown in Figure 1.

All active compounds except the baits were added to the source databases, and search calculations were carried out separately for each class. The resulting distributions of Tc values of all database compounds and known active molecules were monitored and compared in histogram representations in order to analyze the scoring ranges of inactive and active molecules as well as to determine whether the results of the fingerprint search calculations provided a meaningful basis for compound selection and the identification of diverse hits.

RESULTS AND DISCUSSION

Structurally Heterogeneous Activity Classes. Since we were interested in the detection of remote similarity relationships through fingerprint searching, much emphasis was put on the assembly of compound activity classes with distinct structural diversity among their members (Figure 1), despite similar biological activity. Initially, this was done on the basis of visual inspection. For each of the five activity classes, calculation of Tc matrices subsequently provided a more quantitative confirmation of structural diversity among active compounds. A representative example is shown in Figure 2. Here, 28 of 30 GnRH agonists form three distinct clusters or groups of similar compounds, with very little intercluster similarity, and the remaining two compounds are not similar to any others.

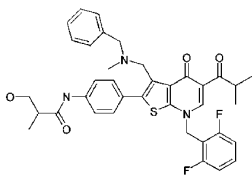
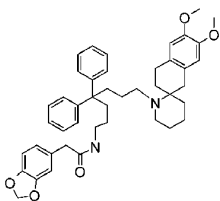
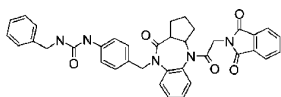
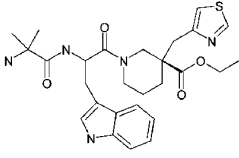
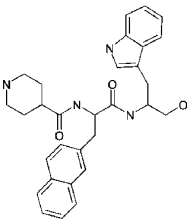
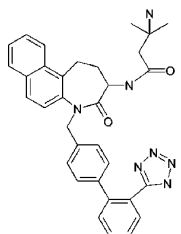
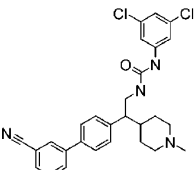
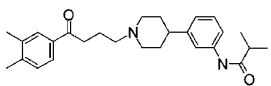
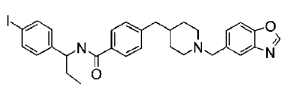
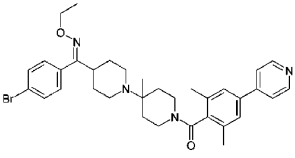
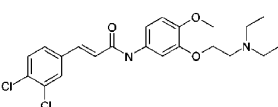
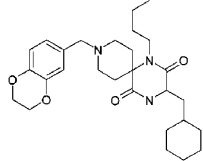
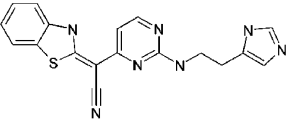
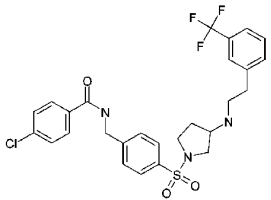
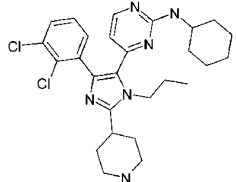
Targets	Most potent compound selected as bait (potency (nm))	Number of reference compounds	
		Examples of reference compounds	
GnRH	 (0.2)	91	
			
GHS	 (0.5)	13	
			
MCH	 (0.4)	31	
			
CCR5	 (0.1)	12	
			
JNK3	 (70)	20	
			

Figure 1. Structures of active compounds. For each activity class, examples of structurally diverse molecules are shown and the total number of test compounds is reported. For fingerprint searching, the most potent compound of each class was used as the bait (or template; shown on the left).

Calculations and Histogram Analysis. In our fingerprint calculations, ACD and MCD, containing ~200 000 and ~900 000 compounds, respectively, were used side-by-side and, as per calculation protocol, between 11 and 90 potential hits were added, dependent on the activity class. Only these actives—but no other database compounds—were considered potential hits. Tc values for all pairwise compound comparisons were separately recorded in a histogram representation for matches between the baits and active compounds

belonging to the same activity class (representing correctly identified hits) and between baits and other ACD or MCD compounds (considered false positives). Figure 3 shows the histogram representations for all search calculations. The potential of successful hit identification and scaffold hopping can be assessed by comparing the Tc ranges of matches between baits and hits and between baits and database compounds (false positives). In a favorable situation, Tc values for bait–hit matches are larger than most of those

#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	1.00	0.65	0.57	0.66	0.72	0.70	0.78	0.72	0.75	0.36	0.43	0.54	0.56	0.58	0.58	0.58	0.60	0.60	0.58	0.60	0.63	0.66	0.61	0.56	0.59	0.53	0.59	0.53	0.59	0.49
2	0.65	1.00	0.77	0.98	0.87	0.82	0.62	0.76	0.77	0.45	0.36	0.57	0.47	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.56	0.57	0.54	0.45	0.46	0.47	0.53	0.47	0.53	0.44
3	0.57	0.77	1.00	0.76	0.75	0.69	0.57	0.64	0.67	0.37	0.32	0.49	0.48	0.44	0.44	0.46	0.46	0.46	0.44	0.46	0.48	0.51	0.47	0.41	0.42	0.47	0.47	0.42	0.47	0.42
4	0.66	0.98	0.76	1.00	0.88	0.84	0.63	0.75	0.78	0.46	0.38	0.57	0.46	0.50	0.50	0.53	0.51	0.51	0.50	0.51	0.55	0.56	0.53	0.47	0.48	0.46	0.52	0.46	0.52	0.44
5	0.72	0.87	0.75	0.88	1.00	0.80	0.67	0.72	0.77	0.43	0.39	0.57	0.50	0.51	0.51	0.57	0.53	0.53	0.51	0.53	0.56	0.58	0.54	0.53	0.54	0.51	0.56	0.51	0.56	0.48
6	0.70	0.82	0.69	0.84	0.80	1.00	0.67	0.79	0.80	0.44	0.34	0.53	0.44	0.46	0.46	0.49	0.47	0.47	0.46	0.47	0.51	0.52	0.49	0.51	0.53	0.48	0.56	0.48	0.56	0.44
7	0.78	0.62	0.57	0.63	0.67	0.67	1.00	0.68	0.75	0.36	0.39	0.51	0.48	0.48	0.48	0.54	0.51	0.51	0.48	0.51	0.53	0.55	0.51	0.56	0.59	0.54	0.57	0.54	0.57	0.63
8	0.72	0.76	0.64	0.75	0.72	0.79	0.68	1.00	0.86	0.42	0.34	0.56	0.47	0.49	0.49	0.50	0.50	0.50	0.49	0.50	0.54	0.55	0.53	0.48	0.51	0.53	0.55	0.53	0.55	0.51
9	0.75	0.77	0.67	0.78	0.77	0.80	0.75	0.86	1.00	0.42	0.35	0.55	0.48	0.49	0.49	0.53	0.51	0.51	0.49	0.51	0.54	0.56	0.52	0.49	0.52	0.51	0.54	0.51	0.54	0.52
10	0.36	0.45	0.37	0.46	0.43	0.44	0.36	0.42	0.42	1.00	0.53	0.58	0.41	0.50	0.50	0.55	0.48	0.48	0.50	0.48	0.47	0.46	0.47	0.48	0.47	0.45	0.43	0.45	0.43	0.41
11	0.43	0.36	0.32	0.38	0.39	0.34	0.39	0.34	0.35	0.53	1.00	0.55	0.49	0.60	0.60	0.60	0.58	0.58	0.60	0.58	0.56	0.54	0.56	0.41	0.38	0.35	0.40	0.35	0.40	0.36
12	0.54	0.57	0.49	0.57	0.57	0.53	0.51	0.56	0.55	0.58	0.55	1.00	0.75	0.91	0.91	0.89	0.88	0.88	0.91	0.88	0.85	0.83	0.85	0.59	0.58	0.59	0.68	0.59	0.68	0.58
13	0.56	0.47	0.48	0.46	0.50	0.44	0.48	0.47	0.48	0.41	0.49	0.75	1.00	0.80	0.80	0.74	0.81	0.81	0.80	0.81	0.76	0.76	0.76	0.53	0.52	0.55	0.63	0.55	0.63	0.54
14	0.58	0.51	0.44	0.50	0.51	0.46	0.48	0.49	0.49	0.50	0.60	0.91	0.80	1.00	1.00	0.87	0.96	0.96	1.00	0.96	0.93	0.90	0.93	0.57	0.55	0.56	0.65	0.56	0.65	0.55
15	0.58	0.51	0.44	0.50	0.51	0.46	0.48	0.49	0.49	0.50	0.60	0.91	0.80	1.00	1.00	0.87	0.96	0.96	1.00	0.96	0.93	0.90	0.93	0.57	0.55	0.56	0.65	0.56	0.65	0.55
16	0.58	0.51	0.46	0.53	0.57	0.49	0.54	0.50	0.53	0.55	0.60	0.89	0.74	0.87	0.87	1.00	0.87	0.87	0.87	0.87	0.82	0.82	0.82	0.64	0.62	0.59	0.67	0.59	0.67	0.57
17	0.60	0.51	0.46	0.51	0.53	0.47	0.51	0.50	0.51	0.48	0.58	0.88	0.81	0.96	0.96	0.87	1.00	1.00	0.96	1.00	0.90	0.93	0.90	0.57	0.56	0.57	0.66	0.57	0.66	0.56
18	0.60	0.51	0.46	0.51	0.53	0.47	0.51	0.50	0.51	0.48	0.58	0.88	0.81	0.96	0.96	0.87	1.00	1.00	0.96	1.00	0.90	0.93	0.90	0.57	0.56	0.57	0.66	0.57	0.66	0.56
19	0.58	0.51	0.44	0.50	0.51	0.46	0.48	0.49	0.49	0.50	0.60	0.91	0.80	1.00	1.00	0.87	0.96	0.96	1.00	0.96	0.93	0.90	0.93	0.57	0.55	0.56	0.65	0.56	0.65	0.55
20	0.60	0.51	0.46	0.51	0.53	0.47	0.51	0.50	0.51	0.48	0.58	0.88	0.81	0.96	0.96	0.87	1.00	1.00	0.96	1.00	0.90	0.93	0.90	0.57	0.56	0.57	0.66	0.57	0.66	0.56
21	0.63	0.56	0.48	0.55	0.56	0.51	0.53	0.54	0.54	0.47	0.56	0.85	0.76	0.93	0.93	0.82	0.90	0.90	0.93	0.90	1.00	0.97	0.97	0.54	0.52	0.53	0.61	0.53	0.61	0.52
22	0.66	0.57	0.51	0.56	0.58	0.52	0.55	0.55	0.56	0.46	0.54	0.83	0.76	0.90	0.90	0.82	0.93	0.93	0.90	0.93	0.97	1.00	0.93	0.54	0.53	0.54	0.62	0.54	0.62	0.53
23	0.61	0.54	0.47	0.53	0.54	0.49	0.51	0.53	0.52	0.47	0.56	0.85	0.76	0.93	0.93	0.82	0.90	0.90	0.93	0.90	0.97	0.93	1.00	0.54	0.52	0.53	0.61	0.53	0.61	0.52
24	0.56	0.45	0.41	0.47	0.53	0.51	0.56	0.48	0.49	0.48	0.41	0.59	0.53	0.57	0.57	0.64	0.57	0.57	0.57	0.57	0.54	0.54	0.54	1.00	0.91	0.89	0.86	0.89	0.86	0.79
25	0.59	0.46	0.42	0.48	0.54	0.53	0.59	0.51	0.52	0.47	0.38	0.58	0.52	0.55	0.55	0.62	0.56	0.56	0.55	0.56	0.52	0.53	0.52	0.91	1.00	0.81	0.79	0.81	0.79	0.73
26	0.53	0.47	0.42	0.46	0.51	0.48	0.54	0.53	0.51	0.45	0.35	0.59	0.55	0.56	0.56	0.59	0.57	0.57	0.56	0.57	0.53	0.54	0.53	0.89	0.81	1.00	0.89	1.00	0.89	0.88
27	0.59	0.53	0.47	0.52	0.56	0.56	0.57	0.55	0.54	0.43	0.40	0.68	0.63	0.65	0.65	0.67	0.66	0.66	0.65	0.66	0.61	0.62	0.61	0.86	0.79	0.89	1.00	0.89	1.00	0.79
28	0.53	0.47	0.42	0.46	0.51	0.48	0.54	0.53	0.51	0.45	0.35	0.59	0.55	0.56	0.56	0.59	0.57	0.57	0.56	0.57	0.53	0.54	0.53	0.89	0.81	1.00	0.89	1.00	0.89	0.88
29	0.59	0.53	0.47	0.52	0.56	0.56	0.57	0.55	0.54	0.43	0.40	0.68	0.63	0.65	0.65	0.67	0.66	0.66	0.65	0.66	0.61	0.62	0.61	0.86	0.79	0.89	1.00	0.89	1.00	0.79
30	0.49	0.44	0.42	0.44	0.48	0.44	0.63	0.51	0.52	0.41	0.36	0.58	0.54	0.55	0.55	0.57	0.56	0.56	0.55	0.56	0.52	0.53	0.52	0.79	0.73	0.88	0.79	0.88	0.79	1.00

Figure 2. Tc matrix. Pairwise comparisons of a subset of 30 GnRH agonists were carried out with MACCS keys, and the Tc values are reported. Tc values greater than 0.7 and 0.8 are shaded light and dark gray, respectively. The bait compound is number 1. With two exceptions, these compounds can be divided into three distinct groups of structurally similar compounds.

calculated for matches between baits and database compounds. In this case, hits could be correctly identified by selecting only a small number of compounds (for example, 100 or fewer), corresponding to low false-positive rates. By contrast, scaffold hopping becomes a more or less hopeless task, at least using fingerprint searching, when many—or the majority—of database compounds are matched at Tc values where true similarity relationships are detected. Then, one would need to select thousands of database molecules, or even more, to obtain compound sets that contain hits.

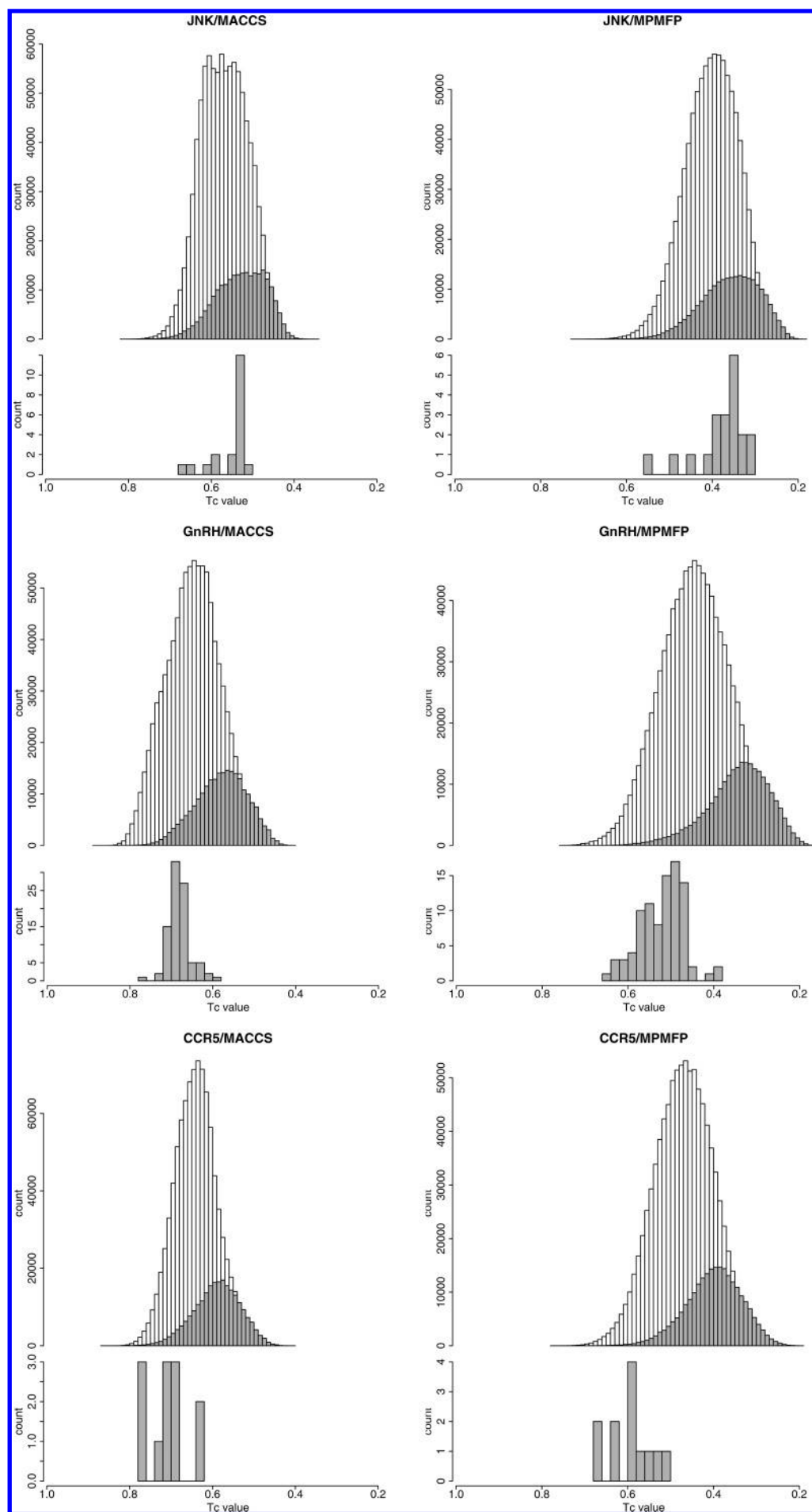
Relevant Tc Threshold Values and Intervals. For fingerprint searching, the choice of similarity threshold values is a critical factor and much debated in the literature. For example, early investigations using Unity fingerprints suggested that a Tc value of 0.85 would indicate at least an 80% chance that two molecules have similar activity,^{5,44} and this value has often been cited in the literature as a similarity threshold, irrespective of the fingerprints applied. On the other hand, more recent studies using Daylight fingerprints suggested that a molecule matching an active template at a Tc value of 0.85 or greater only has a 30% chance to be active.⁶ These and other findings² indicate that there are no general rules and that it is more or less always required to optimize search parameters and similarity threshold values for selected or newly introduced fingerprints.^{23,45} For example, in a comparison of a number of structural fragment-based fingerprints, we found that Tc optima for hit identification generally differed.²⁵ In this study, MACCS had a Tc optimum of ~0.8 and MPMFP, which performed overall best, was shown to detect a number of remote similarity relationships at Tc threshold values between 0.65 and 0.70.²⁵

Recognition of diverse compounds can, in principle, occur over the Tc value range defined by matching the bait with

the structurally most and least similar hits. The Tc distributions of matches between baits and hits in Figure 3 reveal that this interval can span about 20% or even more of the total Tc range. In the case of MPMFP, the least similar hits consistently produced Tc values between 0.3 and 0.5, which represents so limited a structural similarity that one would not expect to correctly identify these bait–hit pairings; one would not consider Tc threshold values within this low range in the search for active compounds.^{2,4,6} This further illustrates the degree of structural diversity within the compound activity classes studied here.

General Trends. Prior to focusing on specific activity classes, a few general conclusions can be drawn from the data presented in Figure 3. With one exception (MCH, where Tc value distributions for hits were similar), MPMFP produced a larger Tc spread for active compounds than MACCS, and MPMFP generally produced broader Tc distributions for (matches between baits and) database compounds. These findings indicate that MPMFP is a more sensitive fingerprint than MACCS, as a result of the addition of binary-encoded molecular property descriptors, consistent with earlier observations.²⁵

Other general trends relate to the databases used as background for similarity searching. Irrespective of the fingerprints, the modes of the Gaussian-like Tc distributions for matches between baits and database compounds are systematically shifted toward lower Tc values for ACD than for MCD. This can be explained by the fact that MCD compounds are generally more relevant for medicinal chemistry purposes than ACD molecules, more druglike, and, therefore, also more similar overall to the leads or drug candidates used as baits. Furthermore, the histograms reveal that, for all activity classes, at least a few diverse hits could be selected together with small numbers of database com-



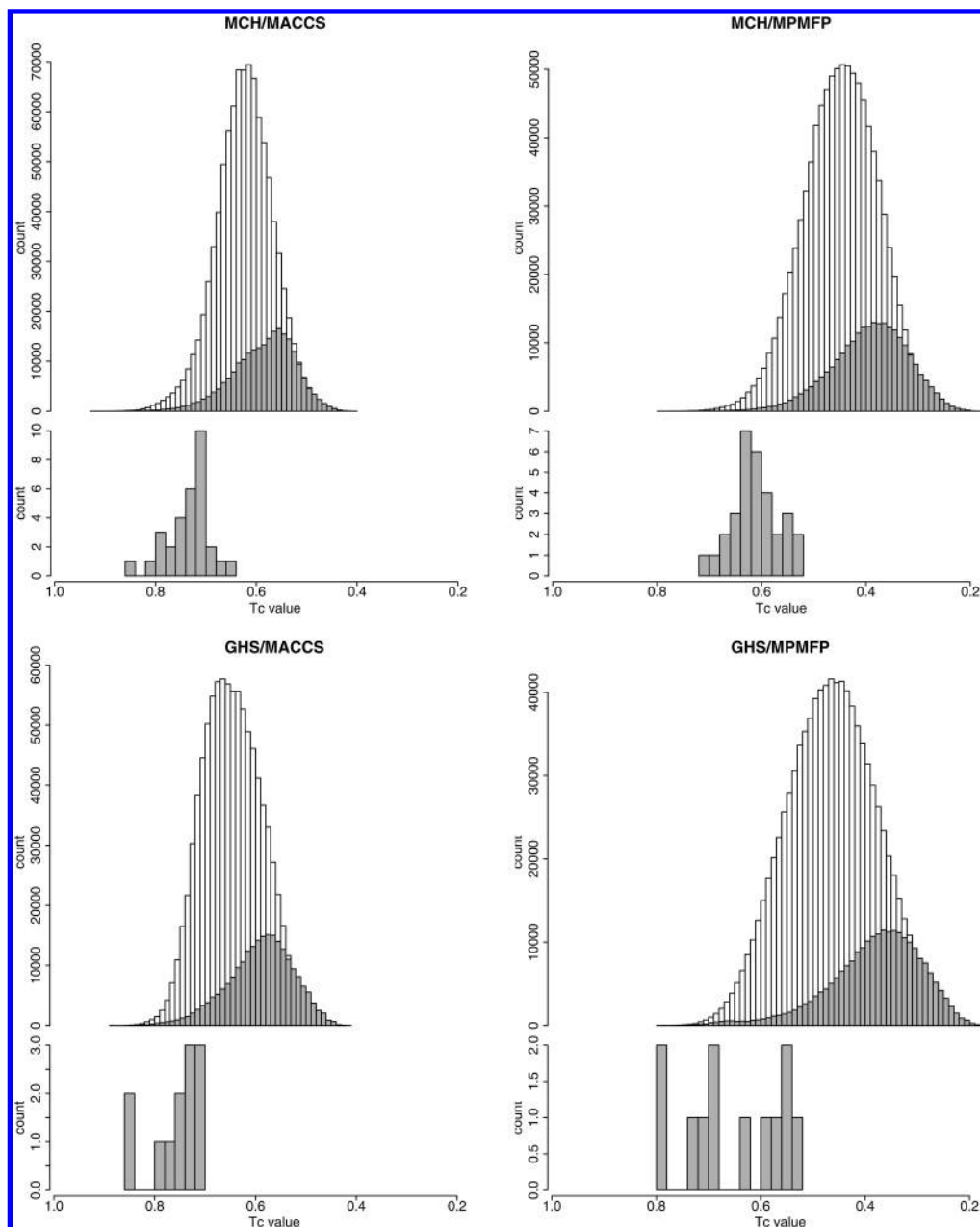


Figure 3. Fingerprint search results. For each compound class, the results of the similarity search calculations are shown in corresponding representations. Each panel is labeled with the activity class studied and the fingerprint used. The upper graph reports the Tc distribution of database compounds obtained for the bait molecule. Open and gray bars represent the MCD and ACD database, respectively. The lower graph reports the number of active compounds detected at a given Tc value and delineates the Tc interval where scaffold hopping occurs.

pounds when ACD was used as the source database. On a first glance, this is a promising result. However, it is largely due to the fact that ACD compounds systematically differ, in fingerprint space, from the more druglike activity classes and MCD compounds. Therefore, using the ACD as a background database for benchmark purposes artificially biases the calculations toward the correct identification of active compounds. For hit identification, one is, of course, much more interested in compounds with high relevance for medicinal chemistry. Thus, MCD provides a more realistic background for the evaluation of scaffold hopping among bioactive molecules. However, its use also increases the difficulty of the search problem because MCD compounds match baits with systematically higher Tc values than ACD molecules, thereby increasing the tendency to produce false positives. For these reasons, the following discussion of

differences between activity classes, as revealed in Figure 3, focuses on MCD-based search calculations.

Activity-Class-Specific Observations. Activity class JNK contained very diverse compounds, as indicated by the significant Tc spread of bait–hit matches, especially when MPMFP was used. In this case, the most similar active compounds matched the bait at low Tc values (<0.6). With MPMFP, only one hit could be identified when selecting ~ 200 MDC compounds. Identifying any other hit with either MACCS or MPMFP would have required the selection of at least 1000 to 2000 database compounds. Strikingly, the majority of MCD compounds matched the bait at least as well, if not better, than JNK inhibitors. Considering the fact that the mode region of the Tc distribution of database compounds corresponds to random matches (no similarity), bait–hit matches within the same Tc range have no predic-

tive value. Thus, most active compounds had completely disappeared in the background Tc noise, making JNK a test case with very limited hit identification and scaffold hopping potential, given the fingerprint calculations carried out here.

The results for the second example in Figure 3, GnRH, were slightly improved for MPMFP, whereas MACCS calculations had very little predictive value. With MPMFP, four agonists matched the bait at a Tc value of ~ 0.65 . Only a few hundred of the 900 000 MCD compounds also matched the bait at this Tc level. However, the majority of GnRH compounds produced a Tc value of ~ 0.5 , which was close to the mode of the database distribution. Thus, while hits could be identified when screening MCD with MPMFP, these calculations had no significant scaffold hopping potential.

A better picture emerged in the case of CCR5. Here, both fingerprints were more sensitive to bait-hit matches and three to four hits could be identified in both cases together with fewer than 100 MDC compounds. Tc distributions of hits were clearly shifted toward higher values relative to the distributions of MDC compounds. Using MPMFP, two different active structures were successfully identified.

MCH was found to be one of the two successful cases, for both fingerprints, although results produced by MPMFP searching were superior. Within the Tc interval [0.60, 0.70], 19 structurally diverse hits were detected and the selection of ~ 100 MDC compounds yielded six of these hits. The other successful case was GHS, where almost no database molecules matched the highest Tc values of bait-hit matches produced with both fingerprints. With MPMFP, five hits belonging to three structure types were identified together with fewer than ~ 100 MDC compounds. Clearly, practical virtual screening applications would have succeeded in both cases, at least when using MPMFP, and benchmark calculations using sets of known MCH or GHS actives added to an MCD-like source database would have assigned a high probability of success to a virtual screening campaign for novel hits having similar activity, since diverse hit-bait relationships were recognized with high sensitivity and false-positive rates could be effectively controlled.

In summary, of the five test cases investigated here, we would consider JNK and GnRH to be rather difficult, if not impossible, MPMFP search tasks, CCR5 an intermediate one, and both MCH and GHS promising ones. The results clearly demonstrate that different compound classes often respond very differently to selected descriptors and fingerprints. In this study, many structurally diverse hits were recognized at Tc values around 0.6 for MFP (and 0.7 for MACCS), and the number of other database compounds simultaneously detected ultimately decided the success or failure of the similarity search calculations.

CONCLUSIONS

In this study, we have investigated fingerprint search calculations to detect diverse hits in large databases from a rather principal point of view. The strong compound class dependence of the results we obtained is striking, as we observed—among the few cases studied here—very different virtual screening situations, ranging from essentially impossible to rather promising ones, at least with respect to the fingerprints used here. These findings emphasize the need

to investigate virtual screening problems on a case-by-case basis. For such purposes, the type of histogram analysis applied here can be very helpful. For example, when series of bait compounds are available for a virtual screening campaign, Tc distributions among baits and database compounds can easily be compared in pilot calculations to estimate the likelihood that true similarity relationships can be detected in the presence of only small numbers of false positives, given available compound databases and search tools. Moreover, if structurally diverse bait sets are available, Tc values or ranges can be estimated where scaffold hopping is likely to occur. Finally, the comparison of Tc value distributions for known active and database compounds makes it generally possible to select small subsets of source databases for further studies. Thus, histogram analysis, as described herein, might not only be useful as a diagnostic but also as a compound preselection tool.

ACKNOWLEDGMENT

The authors thank Ling Xue for many discussions and Britta von der Gönna for help in preparing the manuscript.

REFERENCES AND NOTES

- (1) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
- (2) Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882–894.
- (3) Johnson, M.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; John Wiley & Sons: New York, 1990.
- (4) Patterson, D. E.; Cramer, R. D., III; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behavior: A useful concept for validation of “molecular diversity” descriptors. *J. Med. Chem.* **1996**, *39*, 3049–3059.
- (5) Matter, H. Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional descriptors. *J. Med. Chem.* **1997**, *40*, 1219–1229.
- (6) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45*, 4350–4358.
- (7) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discovery Today* **2002**, *7*, 903–911.
- (8) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity metrics for ligands reflecting the similarity of target proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391–405.
- (9) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.
- (10) Shanmugasundaram, V.; Maggiora, G. M.; Lajiness, M. S. Hit directed nearest-neighbor searching. *J. Med. Chem.* **2005**, *48*, 240–248.
- (11) Shemetulskis, N. E.; Weininger, D.; Blankley, C. J.; Yang, J. J.; Humblet, C. Stigmata: an algorithm to determine structural commonalities in diverse datasets. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 862–871.
- (12) Xue, L.; Stahura, F. L.; Godden, J. W.; Bajorath, J. Fingerprint scaling increases the probability of identifying molecules with similar activity in virtual screening calculations. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 746–753.
- (13) Xue, L.; Godden, J. W.; Stahura, F. L.; Bajorath, J. Profile scaling increases the similarity search performance of molecular fingerprints containing numerical descriptors and structural keys. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1218–1225.
- (14) Wang, R.; Wang, S. How does consensus scoring work for virtual library screening? An idealized computer experiment. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1422–1426.
- (15) Holliday, J. D.; Hu, C.-Y.; Willett, P. Grouping of coefficients for the calculation of intermolecular similarity and dissimilarity using 2D fragment bit strings. *Comb. Chem. High Throughput Screening* **2002**, *5*, 155–166.
- (16) Salim, N.; Holliday, J. D.; Willett, P. Combination of fingerprint-based similarity coefficients using data fusion. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 435–442.

- (17) Edgar, S. J.; Holliday, J. D.; Willett, P. Effectiveness of retrieval in similarity searches of chemical databases: a review of performance measures. *J. Mol. Graph. Model.* **2000**, *18*, 343–357.
- (18) Xue, L.; Godden, J. W.; Stahura, F. L.; Bajorath, J. Similarity search profiles as a diagnostic tool for the analysis of virtual screening calculations. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1275–1281.
- (19) Xue, L.; Godden, J. W.; Stahura, F. L.; Bajorath, J. Similarity search profiling reveals the effects of fingerprint scaling in virtual screening. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2032–2039.
- (20) Cramer, R. D.; Jilek, R. J.; Guessregen, S.; Clark, S. J.; Wendt, B.; Clark, R. D. “Lead hopping”. Validation of topomer similarity as a superior predictor of biological activities. *J. Med. Chem.* **2004**, *47*, 6777–6791.
- (21) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. “Scaffold hopping” by topological pharmacophore search: a contribution to virtual screening. *Angew. Chem., Int. Ed.* **1999**, *38*, 2894–2896.
- (22) Merlot, C.; Domine, D.; Cleve, C.; Church, D. J. Chemical substructures in drug discovery. *Drug Discovery Today* **2003**, *8*, 594–602.
- (23) Xue, L.; Godden, J. W.; Bajorath, J. Mini-fingerprints for virtual screening: design principles and generation of novel prototypes based on information theory. *SAR QSAR Environ. Res.* **2003**, *14*, 27–40.
- (24) MACCS keys. MDL Information Systems Inc.: San Leandro, CA. <http://www.mdll.com>.
- (25) Xue, L.; Godden, J. W.; Stahura, F. L.; Bajorath, J. Design and evaluation of a molecular fingerprint involving the transformation of property descriptor values into a binary classification scheme. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1151–1157.
- (26) MPMFP; available through SVL Exchange, <http://svl.chemcomp.com>.
- (27) *Molecular Operating Environment*; Chemical Computing Group Inc.: Montreal, Quebec, Canada, 2005. <http://www.chemcomp.com>.
- (28) *Available Chemicals Directory*; MDL Information Systems Inc.: San Leandro, CA. <http://www.mdll.com>.
- (29) Millar, R. P.; Zhu, Y.-F.; Chen, C.; Struthers, R. S. Progress towards the development of non-peptide orally-active gonadotropin-releasing hormone (GnRH) antagonists: therapeutic implications. *Br. Med. Bull.* **2000**, *56*, 761–772.
- (30) Walsh, T. F.; Toupence, R. B.; Young, J. R.; Huang, S. X.; Ujjainwalla, F.; DeVita, R. J.; Goulet, M. T.; Wyvratt, M. J.; Fisher, M. H.; Lo, J. L.; Ren, N.; Yudkovitz, J. B.; Yang, Y. T.; Cheng, K.; Smith, R. G. Potent antagonists of gonadotropin releasing hormone receptors derived from quinolone-6-carboxamides. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 443–447.
- (31) Ashton, W. T.; Sisco, R. M.; Yang, Y. T.; Lo, J. L.; Yudkovitz, J. B.; Cheng, K.; Goulet, M. T. Potent nonpeptide GnRH receptor antagonists derived from substituted indole-5-carboxamides and -acetamides bearing a pyridine side-chain terminus. *Bioorg. Med. Chem. Lett.* **2001**, *11*, 1727–1731.
- (32) Ashton, W. T.; Sisco, R. M.; Kieczkowski, G. R.; Yang, Y. T.; Yudkovitz, J. B.; Cui, J.; Mount, G. R.; Ren, R. N.; Wu, T.; Shen, X.; Lyons, K. A.; Mao, A. H.; Arlin, J. R.; Karanam, B. V.; Vincent, S. H.; Cheng, K.; Goulet, M. T. Orally bioavailable, indole-based nonpeptide GnRH receptor antagonists with high potency and functional activity. *Bioorg. Med. Chem. Lett.* **2001**, *11*, 2597–2602.
- (33) GnRH compounds were taken from the following patents: JP9169735, JP9169767, JP1036373, JP11315079, JP200095767, JP2000219690, WO0129044, WO0069859, WO0155119, WO2000069859, JP0812650, JP08253447, JP0948777, JP10500402.
- (34) Nargund, R. P.; Patchett, A. A.; Bach, M. A.; Murphy, M. G.; Smith, R. G. Peptidomimetic GH secretagogues – design considerations and therapeutic potential. *J. Med. Chem.* **1998**, *41*, 3103–3127.
- (35) Hansen, T. K.; Ankersen, M.; Raun, K.; Hansen, B. S. Highly potent growth hormone secretagogues: hybrids of NN703 and ipamorelin. *Bioorg. Med. Chem. Lett.* **2001**, *11*, 1915–8.
- (36) GHS compounds were taken from the following patents: WO0156592, WO9819699, WO9908697, WO0260878, WO9513069, WO0217918, WO9909991, EP0999220.
- (37) Clark, D. E.; Higgs, C.; Wren, S. P.; Dyke, H. J.; Wong, M.; Norman, D.; Lockey, P. M.; Roach, A. G. A virtual screening approach to finding novel and potent antagonists at the melanin-concentrating hormone 1 receptor. *J. Med. Chem.* **2004**, *47*, 3962–3971.
- (38) GnRH compounds were taken from the following patents: WO0204433, WO03004027, WO02057233, WO02076947, WO03045918, WO03047568, WO02076929, WO02083134, WO03059289, WO03060475, WO0206245, WO02010146, WO03033476, WO033480, WO02051809, WO03035055, WO03045920, WO03045313, WO03028641, WO03070244, WO03015769.
- (39) Debnath, A. K. Generation of predictive pharmacophore models for CCR5 antagonists: study with piperidine- and piperazine-based compounds as a new class of HIV-1 entry inhibitors. *J. Med. Chem.* **2003**, *46*, 4501–4515.
- (40) Seibert, C.; Sakmar, T. P. Small molecule antagonists of CCR5 and CXCR4: a promising new class of anti-HIV-1 drugs. *Curr. Pharm. Design* **2004**, *10*, 2041–2062.
- (41) Manning, A. M.; Davis, R. J. Targeting JNK for therapeutic benefit: from junk to gold? *Nat. Rev. Drug Discovery* **2003**, *2*, 554–565.
- (42) Scapin, G.; Patel, S. B.; Lisnock, J.; Becker, J. W.; LoGrasso, P. V. The structure of JNK3 in complex with small molecule inhibitors: structural basis for potency and selectivity. *Chem. Biol.* **2003**, *10*, 705–712.
- (43) JNK compounds were taken from the following patents: WO0147920, WO0123379, WO0123378, WO0123382.
- (44) Brown, R. D.; Martin, Y. C. An evaluation of structural descriptors and clustering methods for use in diversity selection. *SAR QSAR Environ. Res.* **1998**, *9*, 23–39.
- (45) Bajorath, J. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 233–245.

CI050276W