Protocols for Bridging the Peptide to Nonpeptide Gap in Topological Similarity Searches

Robert P. Sheridan,* Suresh B. Singh, Eugene M. Fluder, and Simon K. Kearsley

Department of Molecular Systems, RY50SW-100 Merck Research Laboratories, P.O. Box 2000,
Rahway, New Jersey 07065

Received February 16, 2001

Similarity searches based on chemical descriptors have proven extremely useful in aiding large-scale drug screening. Typically an investigator starts with a "probe", a drug-like molecule with an interesting biological activity, and searches a database to find similar compounds. In some projects, however, the only known actives are peptides, and the investigator needs to identify drug-like actives. 3D similarity methods are able to help in this endeavor but suffer from the necessity of having to specify the active conformation of the probe, something that is not always possible at the beginning of a project. Also, 3D methods are slow and are complicated by the need to generate low-energy conformations. In contrast, topological methods are relatively rapid and do not depend on conformation. However, unmodified topological similarity methods, given a peptide probe, will preferentially select other peptides from a database. In this paper we show some simple protocols that, if used with a standard topological similarity search method, are sufficient to select nonpeptide actives given a peptide probe. We demonstrate these protocols by using 10 peptide-like probes to select appropriate nonpeptide actives from the MDDR database.

INTRODUCTION

Similarity searches are a standard tool for drug discovery.[1,2] The idea behind such searches is that, given a compound with an interesting biological activity, compounds that are "similar" to it in structure are likely to have a similar activity. In practice, an investigator provides a chemical structure as a "probe," searches over a database of sample-available compounds, and finds those database entries that are most similar. These are then submitted for testing. A feature of similarity searching is that the investigator need not specify which parts of a molecule are important for activity. Similarity searching can be done on the basis of topological or 3D structure. Topological similarity searches, especially those based on comparing lists of precomputed substructure descriptors, are computationally very inexpensive.

In most cases the probe is a drug-like molecule. Many times, however, an investigator has only peptide actives. Since peptides generally exhibit poor oral bioavailability, the investigator has to identify nonpeptides with the same activity. This has not been possible in the past with topological similarity searches; peptide probes tend to retrieve other peptides as most similar. Therefore, investigators have used 3D methods,[3−6] which ignore bond topology in favor of atomic coordinates. 3D methods have a number of disadvantages, not the least of which is that it is necessary to specify the active conformation of the probe or the atomic-level structure of the relevant receptor, something that is rarely possible at the beginning of a project. Another disadvantage is that 3D methods are relatively slow and also are limited by the complications of generating and selecting low-energy conformations for the database entries.

We have investigated, using examples from the MDDR database, several simple modifications of our conventional topological similarity protocol to bridge the peptide to nonpeptide gap. These modifications are able to significantly increase the frequency of nonpeptide actives at the beginning of a list sorted by decreasing similarity to a peptide probe.

METHODS

Definitions of Descriptors. Here we will use atom pair and the topological torsion descriptors. Atom pairs are substructures of the form

$$AT_i \text{ --} AT_j \text{--(distance)}$$

where AT_i is the atom type of i and (distance) is the distance in bonds between atom i and atom j along the shortest path. The topological torsion is of the form

$$AT_i \text{ -- } AT_j \text{ -- } AT_k \text{ -- } AT_l$$

where i, j, k, and l are consecutively bonded atoms. For the "regular" atom pair (AP), originally described by Carhart et al.,[7] and the regular topological torsion (TT), originally described by Nilakantan et al.[8], AT_i contains information about the element, number of non-hydrogen neighbors, and number of π electrons for atom i. A previous publication from our laboratory[9] introduced the binding point pair (BP) and binding point torsion (BT), wherein AT_i is one of seven physiochemical types (1 = cation, 2 = anion, 3 = donor, 4 = acceptor, 5 = polar, 6 = hydrophobe, 7 = other). Rules for assigning these types, given only the connection table of a molecule, have been published.[10] Only non-hydrogen atoms are considered. Reference 9 shows an example of a molecule parsed into these four descriptors.

Combination Descriptors. Also in ref 9, we explored the concept of combination descriptors wherein the similarity

* Corresponding author phone: (732)594-3859; fax: (732)594-4224; e-mail: sheridan@merck.com.

10.1021/ci0100144 CCC: $20.00
Published on Web 08/01/2001

score of a database molecule with the probe is the mean of the similarities with each individual descriptor. For example, consider the APTT descriptor:

$$\mathrm{Sim(APTT)} = 0.5\mathrm{Sim(AP)} + 0.5\mathrm{Sim(TT)}$$

The idea of combination descriptors is that combining a specific descriptor (e.g. AP) with a fuzzy descriptor (e.g. BP) can often get results better than with either alone.[9]

Definition of Similarity. In all cases we will use the Dice definition of similarity. The similarity of molecules A and B is

$$\mathrm{Sim}_{AB} = \frac{\sum_j \min(d_{jA}, d_{jB})}{0.5[\sum_j d_{jA} + \sum_j d_{jB}]} \tag{1}$$

where d_{jA} is the count of descriptor j in molecule A. The index j goes over the union of unique descriptors in A and B. Sim_{AB} ranges from 0.0 (nothing in common) to 1.0 (identity). Two other popular definitions of similarity are Tanimoto and cosine. We have found that Dice usually is more effective in finding actives than cosine.[11] For any given probe, Tanimoto is monotonic with Dice.[2]

Sorting of Scores. In a search, the similarity of each database entry to the probe is calculated using a descriptor or combination descriptor. The database entries are sorted from high to low similarity. Ranks are then assigned: the molecule with the highest similarity is rank 1, the next highest rank 2, etc. We use only the ranks of the compounds in this study, since the distribution of absolute similarities varies from one descriptor to another and from one protocol to another.

Measures of Merit for Similarity Searches. Our usual measures of merit are based on a retrospective screening experiment. Imagine a database of N entries. The entries are "tested" in order of increasing rank. The cumulative number of actives found is monitored as a function of candidates tested. This produces an accumulation curve (sometimes called a cumulative recall curve), which is typically hyperbolic if the front of the sorted database is enriched in actives.[9,12] In the past[9] we have used "initial enhancement" (IE), which is the ratio of the observed number of actives at n_{cutoff}, some arbitrary number of compounds tested ($n_{\mathrm{cutoff}} \ll$ N), vs the number expected if the actives were randomly distributed in the list ($n_{\mathrm{actives}} \cdot n_{\mathrm{cutoff}}/N$). Generally we set $n_{\mathrm{cutoff}} = 300$. However, if the number of actives is small, this measure is not very robust because moving one or two compounds across the arbitrary boundary makes a large difference in IE. Therefore, for this work we propose a smoother function to count the number of actives at the beginning of the sorted list

$$S = \sum_i^{\mathrm{actives}} \exp(-\mathrm{rank}(i)/a) \tag{2}$$

where $a \ll N$. If active i is near the front of the list, it will have a weight approaching 1. The weight of an active falls off smoothly as its rank increases. The robust initial enhancement will be

$$\mathrm{RIE} = S/\langle S\rangle \tag{3}$$

where $\langle S\rangle$ represents the mean S calculated from 1000 trials where the ranks of the actives are randomly reassigned a value from 1 to N while ensuring no two actives have the same rank. This method also allows us to monitor in what fraction of the randomized trials the value of S is at least as large as the observed one. RIE = 1 indicates no enhancement over chance.

It happens that the area under the curve for $\exp(-i/a)$ is the same as for a box function where $f(i) = 1$ for $0 \le i \le a$ and 0 for $i > a$. Thus, to be most consistent with IE and $n_{\mathrm{cutoff}} = 300$, we take $a = 300$, and we will use this value throughout. RIE is highly correlated with IE when the number of actives is large ($R^2 > 0.90$). The best fit line is RIE = 0.80 IE.

Database Used in This Study. To validate the search methods we need to have a database of molecules for which we know the biological activities. For this purpose, we use the MDDR (MDL Drug Data Report) database[13] version 98.1 which contains ~82 000 compounds. Most structures have one or more key words in the "therapeutic category" field. We will assume that a molecule is active as an oxytocin antagonist, for instance, if it contains the key word "oxytocin antagonist" in this field. There are some unavoidable limitations to using patent databases such as MDDR. Since most compounds have been tested in only one or two therapeutic areas, one cannot assume that a compound without a particular key word is inactive in the corresponding therapeutic area. Thus there are probably many false inactives. The opposite problem is that for some key words not all actives work by the same mechanism as the probe (for instance by binding to the same receptor site), and we should not necessarily expect all actives to resemble the probe. Thus there are also some false actives. Despite these complications, comparisons between similarity methods should be valid, because for any given probe the number of false actives and false inactives is the same for all methods.

Definition of Nonpeptide. There are many possible ways to define "peptide", but for our purposes we will consider a molecule a peptide if it includes the substructure

$$\mathrm{N\text{-}Csp^3\text{-}C(=O)\text{-}N\text{-}Csp^3\text{-}C(=O)}$$

that is, if there are two adjacent alpha-amino acids. "Nonpeptides" are everything else. Using a substructure definition has the advantage that one can unambiguously classify any compound. However, there is still room for argument whether particular nonpeptides maintain a lot of "peptide flavor" or whether particular peptides are modified so much that they are drug-like.

Choice of Example Probes for Similarity Searches. The probes are shown in Figure 1, and Table 1 shows the therapeutic keywords used to define actives. Only nonpeptide actives are considered. The probes and the corresponding therapeutic category were arbitrarily selected such that the following was true:

(1) The probe was a heptapeptide or smaller.

(2) Compounds in the same therapeutic category as the probe were fairly numerous, and several chemical classes of nondipeptide actives were present.

(3) The therapeutic category was fairly specific, so that most of the molecules probably work by the same mechanism.

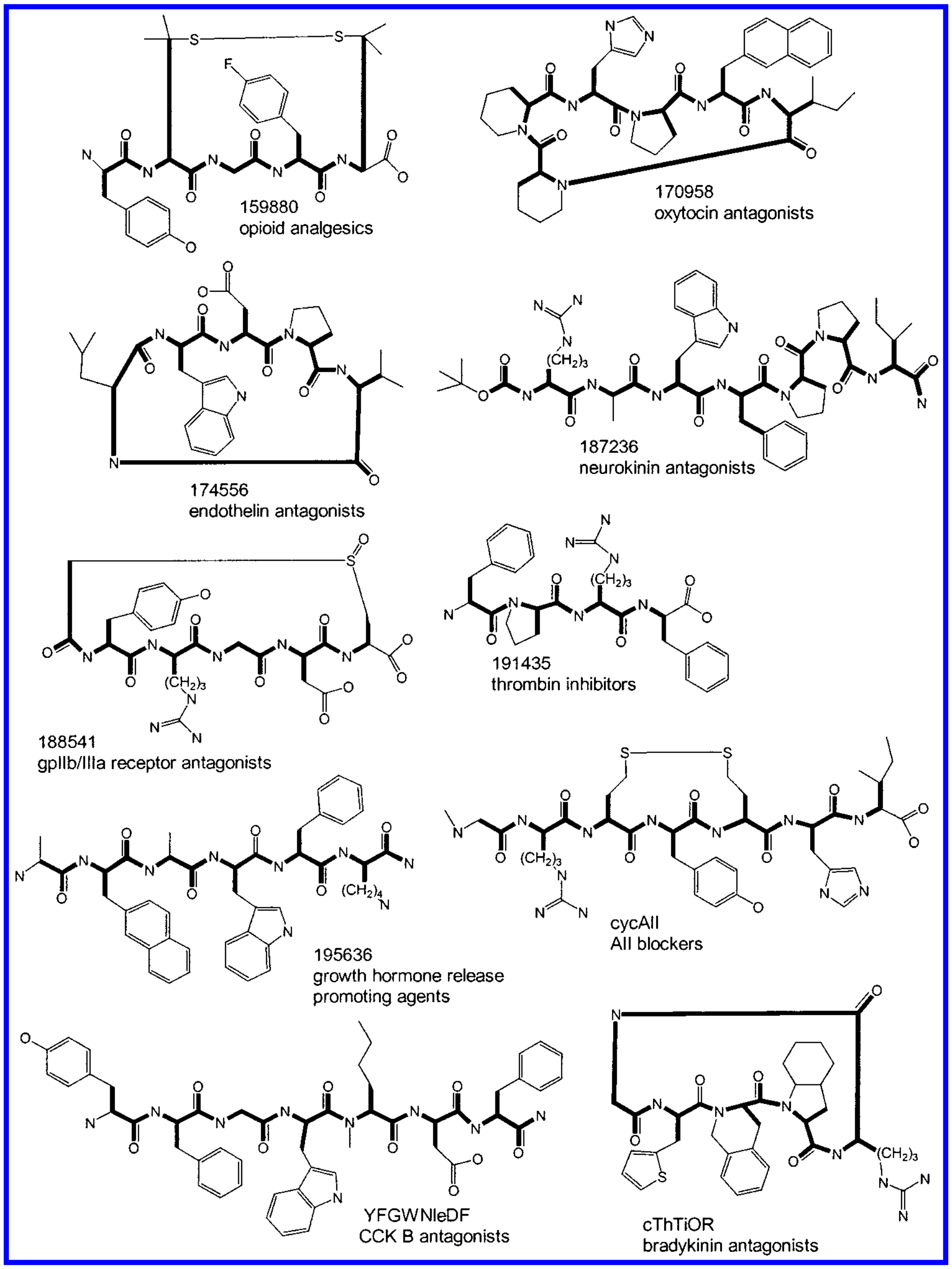


Figure 1. Peptide-like probes used for this study. Atoms connected by bold bonds are those eliminated in the **backbone** and **backbonegone** protocols.

Seven of the probes are from the MDDR. Three are from the literature. The probe cycAII is from Spear et al.[14] YFGWNleDF is from Nikiforovich et al.[15] and cGThTiOR from Thurieau et al.[16] It happens that the majority of activities that meet the criteria are mediated by G-protein coupled receptors (GPCRs). Generally the actives contain more atoms than the average nonpeptide compound from the MDDR, but the actives are also much smaller than the probe.

The Protocols. Peptides differ from drug-like molecules in a number of ways:

1. They have a repeating backbone structure (N−C−C=O)n.

2. They generally have more atoms.

3. They generally have longer through-bond distances.

Because the backbone atoms are a large fraction of the total atoms, a similarity search on any peptide probe will be heavily biased toward peptides, regardless of the biological activity of the probe. The protocols proposed here are aimed toward eliminating the bias:

1. Database filtering. This protocol is the most obvious: simply eliminate peptides from the database. The effect of this is best explained by looking at an example. Figure 2A shows the distribution of AP similarities of peptides and nonpeptides from the MDDR to the probe 188541. Clearly the peptides have a higher mean similarity (0.39 ± 0.10) than the nonpeptides (0.18 ± 0.08). However, the similarities of the appropriate actives, nonpeptide gpIIb/IIIa antagonists, are somewhat higher (0.29 ± 0.06) than the total set of nonpeptides. The accumulation curve for all molecules ("full database") in Figure 2B is nicely hyperbolic. However, because peptides are dominating the front of the list, the first active occurs at rank 986, and the RIE is 0.0, that is, there are fewer nonpeptide actives at the front of the list than expected by chance. Eliminating the ~6700 peptides in the

Table 1. Probes and Activity Keywords Used in This Study

name of probe	activity keywords from MDDR	no. of nonpeptide actives	mean ± SD[a] non-hydrogens in actives	non-hydrogens in probe
159880	opioid analgesics opioid agonist kappa agonist delta agonist mu agonist	735	27 ± 5	45
170958	oxytocin antagonist	159	36 ± 6	56
174556	endothelin antagonist	488	35 ± 7	44
187236	neurokinin antagonist	105	38 ± 5	71
188541	gpIIb/IIIa receptor antagonist	795	30 ± 5	46
191435	thrombin inhibitors	194	36 ± 15	45
195636	growth hormone releasing agent	93	41 ± 4	60
cycAII	angiotensin II blocker angiotensin AT1 antagonist angiotensin AT2 antagonist	2216	36 ± 6	68
YFGWNleDF	CCK B antagonist	177	39 ± 6	70
cGThTiOR	bradykinin antagonist	100	42 ± 7	48

[a] Compared to 30 ± 11 for all MDDR nonpeptides.

MDDR shifts the curve ("filtered database") toward the left. The first active occurs at rank 41, and the RIE is a respectable 7.2. It should be emphasized that for there to be a significant enhancement of actives at the front of the list of nonpeptides, it is not necessary that the absolute similarity of the actives be high, only that some actives are on the average more similar to the probe than the bulk of the inactives. Since it is guaranteed to improve the results, all the results reported here, unless stated otherwise, will be for a filtered database. This protocol where, other than the filtering, the similarity method is not modified will be labeled **default**.

2. Setting the frequencies of long-range (in through-bond distance) AP or BP descriptors in the probe to zero. (TT and BT descriptors are not affected.) This has two effects. One is to eliminate the possibility of matching the long-range descriptors in the probe. The other is to reduce the effective size of the probe, thus favoring smaller molecules from the database. We tried three cutoffs for what could be considered "long-range": >5 bonds (**gt5**), >10 bonds (**gt10**), and >15 bonds (**gt15**).

3. Setting the frequencies of all descriptors to 1 or 0 for "present" or "absent". This makes the descriptors behave like a fingerprint method.[17−19] The hope is that the influence of descriptors referring to the backbone, which are repeated in peptides, will be downweighted. This protocol will be called **fingerprint**.

4. Eliminating the backbone atoms. Here we change atoms in the backbone (N−C−C═O) so that descriptors containing them cannot be matched. In our implementation the element type is set to "Na" for AP or TT and the physiochemical type to "9" for BP and BT. The β-carbon is eliminated too if the amino acid is not Ala or Gly. If the end of the peptide backbone is a free carboxylate or amine, those atoms are kept. This is called the **backbone** protocol. An alternative is to set the frequencies of descriptors containing the eliminated atoms to zero. This produces the same numerator in eq 1, but makes the probe look effectively smaller in the denominator. This is called **backbonegone** protocol.

5. How much the database entry resembles the probe more than it resembles an "average peptide". For this, the descriptor centroid[20] for the ∼6700 peptides in the MDDR was calculated and used to rank the database entries on the basis of decreasing similarity to this centroid. The database entry was also ranked in order of decreasing similarity to the probe, as in the **default** protocol. The "disparity score" of a database entry i is

$$\text{disparity score} = -\log[\text{rank}(i, \text{probe})/\text{rank}(i, \text{peptide centroid})] \quad (4)$$

Database entries are sorted by decreasing disparity score. This is called the **rankratio** protocol.

RESULTS

RIEs are shown in Table 2. Note that all RIEs are for a filtered database. If the database is not filtered (data not shown), only the **rankratio** and **backbone** protocols can produce RIEs $\gg 1$. This is expected because **rankratio** selects against peptide-like molecules, and **backbone** nullifies the influence of the backbone. For all other protocols, all RIEs are near zero. Filtering always improves the RIEs, even for the **rankratio** and **backbone** protocols.

The RIE for each protocol/descriptor combination, averaged over all 10 examples, is shown at the end of Table 2 and labeled "mean all probes". This is a semiquantitative indication of how good each protocol/descriptor combination is overall. Generally, individual RIEs of ≥ 3 can be regarded as statistically different from chance at the $P < 0.01$ level.

Several general things are evident from this table:

1. Overall the enhancements are quite modest (2−10) compared to the results from drug-like probes (typically 20−50 from ref 9), although in some individual cases, the RIEs can be quite high (>30).

2. For any given probe, the results are quite sensitive to protocol and to the descriptor. Conversely, which protocol/descriptor combination is best varies from probe to probe in a way not easy to predict beforehand. This observation is not surprising, since similar observations are true for drug-like probes.[9]

3. There are some probes that have RIEs ≥ 5 for nearly all protocol/descriptor combinations (e.g. 191435) and one probe (cycAII) that has RIEs < 1 for all combinations. That

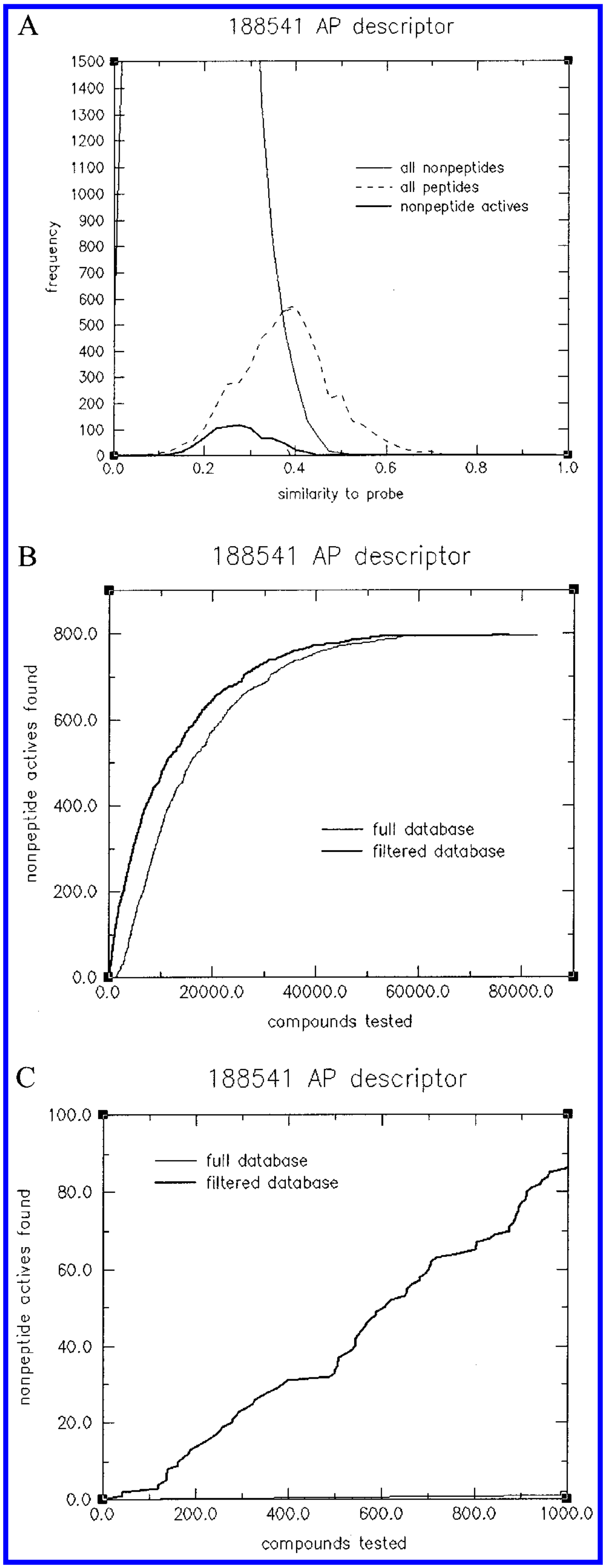


Figure 2. A. Distribution of absolute similarities to 188541 for all peptides in the MDDR, all nonpeptides, and the relevant actives (nonpeptide gpIIb/IIIa receptor antagonists). B. The accumulation curves for the full (all molecules) and filtered database (peptides removed). An accumulation curve monitors the number of nonpeptide actives found as the database is tested in order of decreasing similarity to the probe. C. A closeup of the origin of B.

there is at least one protocol/descriptor combination with an RIE > 5 for most probes is encouraging.

4. Sometimes combination descriptors give higher RIEs than the individual descriptors. This is also observed for drug-like probes.[9]

It is conceivable that the high RIEs are due, not to a specific resemblance of probes and actives, but to the fact that the overall set of actives (mostly GPCR-related and perhaps designed to mimic peptides) are somehow anomalous compared to most drug-like molecules. That is, perhaps the similarity searches are picking up nothing but nonspecific characters such as peptide flavor or large size. To examine this possibility we recalculated the RIE for every probe/descriptor/protocol combination, this time using two artificial lists of pseudoactives. The first list of pseudoactives ("non-dipeptide pseudoactives") was made from 500 randomly selected nondipeptides from the MDDR and used for all probes. (500 is close to the mean size of the 10 active lists.) A second list of pseudoactives ("active-like pseudoactives") was made for each individual probe by randomly selecting 500 molecules from the active lists of the other nine probes. As expected, these control RIEs did not show any systematic effect with probe, protocol, or descriptor and are skewed to lower values. The nondipeptide pseudoactive RIEs are distributed pretty much as expected by chance: 50% of the RIEs are below 1.0 and 99% below 2.1. The active-like pseudoactive RIEs are somewhat higher: 50% of them are below 1.2 and 99% of them are below 5.0, showing that the actives in Table 1 are indeed somewhat anomalous. Thus, the RIEs in Table 2 less than 5 may be due to nonspecific effects, but many of the RIEs are much larger than this and are probably due to specific resemblances.

In comparing protocols we note the following:

1. In individual cases the "gt" protocols can be as good or better than **default** for AP and BP. **gt10** seems the best overall, although for some individual probes the best results are found with **gt5** (e.g., 159880 BP and 170958 BPBT). Perhaps the **gt5** protocol does so well in the 159880 case because the actives are so small (27 ± 5 nonhydrogen atoms) compared to the other sets of actives in this study. However, the actives for 170958 are not particularly small (36 ± 6 atoms).

2. On the average **fingerprint** protocol is a slight improvement over **default** except for the AP and BT descriptors.

3. The **backbone** protocol is sometimes better than **default**, but only with AP or BP descriptors or combination descriptors containing them. The **backbone** protocol with TT and BT descriptors is much worse than **default**. This is not surprising because, once the backbone is removed, there are very few TTs and BTs left in most probes. The **backbonegone** protocol is uniformly worse than **backbone**, again probably because this makes the effective size of the probe much too small.

4. The **rankratio** protocol is somewhat better than **default** only for AP descriptors and the combinations that contain them. However, on inspecting which actives are selected, **rankratio** does not seem to be selecting compounds so different from the other protocols than the extra complexity of doing two separate similarity searches is justified.

Figure 3 shows some of the actives found for selected protocol/descriptor combinations, where the rank ≤ 300. For each drawing the MDDR external registry number and the

Table 2. Robust Initial Enhancements ($a = 300$) for Filtered Database

	AP	TT	BP	BT	APTT	BPBT	TTBT	APTTBPBT	APBP
159880/Opioid Analgesics									
default	1.9	4.8	1.1	3.1	3.5	2.4	3.9	2.8	1.8
gt5	4.9	4.8	8.9	3.1	6.4	3.5	3.9	5.5	4.6
gt10	2.8	4.8	1.0	3.1	5.0	2.5	3.9	3.6	2.0
gt15	1.9	4.8	1.1	3.1	3.7	2.4	3.9	2.9	1.9
fingerprint	2.5	4.8	2.0	6.8	3.4	6.1	6.3	5.7	2.7
backbone	2.4	0.5	0.0	1.3	0.7	1.7	0.6	0.6	1.8
backbonegone	3.2	0.4	0.7	0.7	0.7	1.2	0.5	0.6	3.3
rankratio	3.3	0.8	1.6	3.3	2.6	3.6	1.5	5.5	3.2
170958/Oxytocin Antagonists									
default	0.6	0.0	1.6	31.1	0.1	9.5	3.0	1.2	0.7
gt5	0.0	0.0	14.8	31.1	0.0	56.7	3.0	7.2	3.5
gt10	0.2	0.0	15.3	31.1	0.0	27.6	3.0	2.7	1.4
gt15	0.5	0.0	1.7	31.1	0.0	9.9	3.0	1.2	0.6
fingerprint	2.0	0.0	14.3	21.5	0.2	21.2	5.1	7.7	10.1
backbone	1.5	0.2	0.4	0.8	1.0	4.2	0.1	2.7	3.6
backbonegone	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
rankratio	0.3	4.6	2.5	42.2	0.2	21.6	28.0	14.1	1.3
174556/Endothelin Antagonists									
default	8.8	5.4	11.9	1.5	7.7	5.9	5.0	8.4	11.2
gt5	1.6	5.4	0.7	1.5	5.7	1.2	5.0	5.8	0.8
gt10	9.5	5.4	10.9	1.5	7.9	6.1	5.0	8.8	12.0
gt15	8.8	5.4	11.9	1.5	7.7	5.9	5.0	8.4	11.2
fingerprint	7.2	2.8	9.3	0.2	6.1	2.2	1.0	6.0	12.4
backbone	3.5	1.9	3.6	0.0	2.8	1.7	0.8	3.3	7.3
backbonegone	0.0	0.2	0.0	0.0	0.2	0.0	0.0	0.1	0.0
rankratio	9.3	1.6	13.8	0.6	7.4	9.3	2.0	12.8	14.1
187236/Neurokinin Antagonist									
default	0.0	8.5	0.1	15.3	3.1	4.3	6.1	1.4	0.0
gt5	0.1	8.5	0.1	15.3	6.9	14.3	6.1	4.3	0.0
gt10	0.3	8.5	14.4	15.3	5.0	28.1	6.1	3.3	2.3
gt15	0.3	8.5	1.2	15.3	4.1	11.8	6.1	2.3	0.2
fingerprint	0.1	7.2	0.0	1.9	5.8	2.3	2.8	2.3	0.0
backbone	4.9	11.6	2.2	1.7	15.2	8.7	12.2	16.4	3.0
backbonegone	4.8	4.2	0.8	0.0	8.7	5.3	2.4	12.1	1.9
rankratio	0.6	11.5	0.1	1.1	2.9	0.1	2.7	0.1	0.1
188541/gpIIb/IIIa Antagonists									
default	7.2	2.3	2.2	7.2	3.3	8.1	3.0	4.5	7.9
gt5	6.4	2.3	0.0	7.2	4.4	4.4	3.0	5.9	2.3
gt10	9.3	2.3	0.8	7.2	3.1	6.9	3.0	4.3	7.8
gt15	7.8	2.3	2.3	7.2	3.4	8.2	3.0	4.6	8.5
fingerprint	11.7	4.6	6.2	7.1	7.6	14.0	5.1	10.2	17.7
backbone	8.5	4.5	29.3	0.7	12.5	13.7	2.4	16.0	26.6
backbonegone	13.4	2.2	40.7	0.4	8.8	4.7	1.3	9.1	37.1
rankratio	19.9	5.6	3.3	8.0	16.0	10.2	11.9	16.0	15.1
191435/Thrombin Inhibitors									
default	12.0	7.3	4.7	19.9	10.2	20.6	14.0	15.1	13.9
gt5	9.3	7.3	0.1	19.9	10.8	10.5	14.0	15.5	4.4
gt10	14.1	7.3	5.0	19.9	11.3	24.0	14.0	17.8	17.2
gt15	13.2	7.3	13.1	19.9	10.9	24.3	14.0	16.8	17.5
fingerprint	15.1	24.4	1.1	18.0	26.3	14.1	24.5	24.2	9.8
backbone	18.6	0.1	10.2	1.5	5.9	24.7	0.8	19.1	23.6
backbonegone	14.9	0.0	4.3	0.1	2.0	0.9	0.2	2.2	13.0
rankratio	31.2	5.9	13.3	17.1	14.0	32.7	13.3	29.1	30.8
195636/GH Releasing Agents									
default	13.7	17.3	4.8	0.4	24.6	3.0	8.0	11.1	9.2
gt5	6.0	17.3	0.0	0.4	19.0	0.2	8.0	12.3	4.2
gt10	37.5	17.3	35.4	0.4	29.5	12.7	8.0	21.1	41.3
gt15	20.9	17.3	9.3	0.4	27.4	4.5	8.0	13.1	15.1
fingerprint	14.4	3.6	41.6	0.0	13.5	8.6	1.7	18.3	49.1
backbone	25.1	1.9	2.4	0.4	23.9	4.8	5.0	27.9	23.1
backbonegone	11.2	0.1	0.1	0.0	6.5	0.2	0.5	5.8	4.7
rankratio	5.2	0.7	0.0	0.0	2.1	0.0	0.2	1.5	3.7
cycAII/AII Blockers									
default	0.1	0.1	0.6	0.1	0.1	0.1	0.0	0.1	0.1
gt5	0.1	0.1	0.0	0.1	0.1	0.1	0.0	0.1	0.0
gt10	0.0	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0
gt15	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.1
fingerprint	0.3	0.2	0.1	0.4	0.1	0.0	0.1	0.0	0.2
backbone	1.6	0.0	0.5	0.1	1.0	0.0	0.0	0.2	1.4
backbonegone	0.3	0.0	0.1	0.0	0.1	0.0	0.0	0.0	0.1
rankratio	0.2	0.6	0.1	0.5	0.2	0.1	0.2	0.1	0.1

Table 2 (Continued)

	AP	TT	BP	BT	APTT	BPBT	TTBT	APTTBPBT	APBP
YFGWNleDF/CCKB Antagonists									
default	13.2	1.6	0.7	26.6	5.4	14.4	11.1	11.1	7.4
gt5	4.8	1.6	1.0	26.6	1.9	21.0	11.1	12.3	6.1
gt10	32.9	1.6	40.6	26.6	4.2	28.8	11.1	16.2	47.1
gt15	22.2	1.6	7.8	26.6	6.8	20.6	11.1	14.6	18.0
fingerprint	1.0	11.9	0.0	9.5	10.9	3.3	18.8	12.3	0.0
backbone	0.1	1.4	0.1	0.0	0.1	3.6	0.7	0.2	0.0
backbonegone	0.0	0.9	0.1	0.0	0.1	0.0	0.1	0.0	0.0
rankratio	11.8	1.7	0.2	18.4	2.9	8.6	1.8	5.4	7.4
cGThTiOR/Bradykinin Antagonists									
default	1.9	0.2	17.7	2.5	0.9	11.8	2.5	5.4	12.0
gt5	0.0	0.2	0.0	2.5	0.0	0.2	2.5	0.4	0.0
gt10	0.8	0.2	7.0	2.5	0.5	5.3	2.5	2.9	6.8
gt15	1.9	0.2	15.8	2.5	0.9	10.7	2.5	5.1	11.3
fingerprint	3.4	0.8	10.5	6.4	2.5	23.4	4.4	11.1	12.6
backbone	11.0	0.2	8.5	0.0	1.6	1.7	0.5	4.1	23.6
backbonegone	3.6	0.1	1.1	0.0	1.7	0.1	0.1	2.0	3.5
rankratio	0.3	1.5	12.2	0.1	1.7	4.8	1.4	2.6	9.3
Mean All Probes									
default	5.9	4.8	4.5	10.8	5.9	8.0	5.7	6.1	6.4
gt5	3.3	4.8	2.6	10.8	5.5	11.2	5.7	6.9	2.6
gt10	10.7	4.8	13.0	10.8	6.7	14.2	5.7	8.1	13.8
gt15	7.8	4.8	6.4	10.8	6.5	9.8	5.7	6.9	8.4
fingerprint	5.8	6.0	8.5	7.2	7.6	9.5	9.0	9.8	11.5
backbone	7.7	2.2	5.7	0.7	6.5	6.5	2.3	9.1	10.7
backbonegone	5.1	0.8	4.8	0.1	2.9	1.2	0.5	3.2	6.4
rankratio	8.2	3.5	4.7	9.1	5.0	9.1	6.3	8.7	8.5

rank in the search is given. Where there are close analogues to an active and the analogues also are active and have ranks ≤ 300, the number of such analogues is given. Generally the actives in Figure 3 have fairly low ranks, that is, at least one interesting active would be found quickly in the retrospective assay. Consistent with observations from drug-like probes,[9] changing the protocol/descriptor combination changes which chemical class of actives is most favored.

A trivial result would have been for the molecules in Figure 3 to be very peptide-like, just missing our definition of "peptide" by some slight modification of the backbone or inclusion of beta-amino acids. However, except perhaps for some of the opioid analgesics, most of the actives seem to be satisfyingly drug-like. That said, it should be noted that some peptide character (i.e. amide bonds, perhaps one alpha amino acid) is retained in many of them. A startling exception, where the protocol seems to have made a leap from peptide to completely drug-like molecules, is for 159880 **gt5** BP, wherein an enkephalin-like probe has found morphine-like compounds!

There are three probes, 191435, 195636, and YFGWNleF, which seem to high RIEs for most protocol /descriptor descriptor combinations, and one probe, cycAII, that never has a high RIE. Might this reflect the different peptide flavor of the nonpeptide actives? One measure of the peptide flavor would be a relatively high similarity of the actives to the peptide centroid used for the **rankratio** protocol. However, although the growth hormone release agents (195636) and CCK B antagonists (YFGWNleF) have high similarities to the centroid by the AP or TT descriptor, thrombin inhibitors (191435) are only moderately similar, as are angiotensin blockers (cycAII). Conversely, bradykinin antagonists are very similar to the centroid, but cGThTioR does not seem to have especially high RIEs. There also is no relationship of having high RIEs with the mean number of atoms in the actives. Again we see that the RIEs are not controlled by nonspecific properties of the actives such as size or peptide flavor.

One can retrospectively pick out specific salient features of the peptide in the nonpeptide actives, and this gives us further confidence that there is something specific in the actives that the probe is selecting. For instance, a clear example is the 188541 probe, which contains Arg and Asp; the actives tend to contain a cation at one end of the molecule and an anion at the other. In the 170958 case, one notices the presence of imidazole in the probe and actives. On the other hand, in examples such as YFGWNleDF, what features are common to probe and actives, other than the presence of aromatic groups, is not so obvious.

DISCUSSION

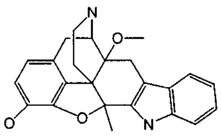
We have demonstrated that it is possible, using a set of simple protocols, for topological similarity methods to find nonpeptide actives given a peptide probe. This is very useful early in a project when one has only peptide actives and too little information to use more sophisticated search methods. The methodology may select molecules with at least some peptide character (e.g. amide bonds), but while perhaps not ideal as drugs, they are drug-like enough to be acceptable as leads for further work.

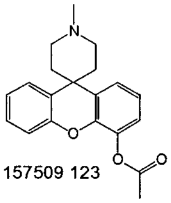
Conventional similarity searches, wherein one has a drug-like probe and is looking in the database for other drug-like entries, are straightforward in that we are usually interested in the database entries most similar to the probe. In the case of a peptide probe, we are trying to find database entries with some characteristics of the probe but need special protocols to exclude molecules that are the most similar, i.e., other peptides.

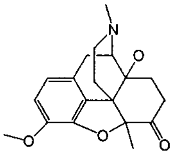
Some of our protocols require modification of the probe descriptors, but none requires that we change the descriptors of the database. This is fortunate. Keeping different versions

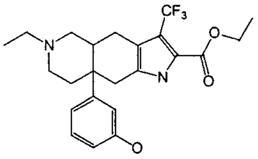
159880/opioid analgesics

gt5 BP


233054 8
+11 analogs

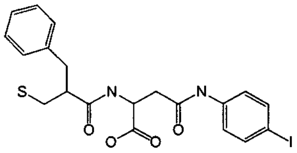

157509 123

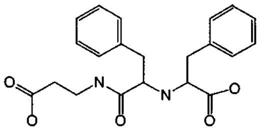

091305 133
+4 analogs

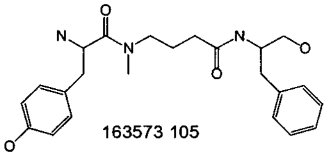


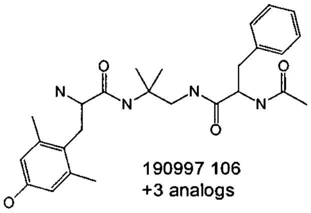
223205 183
+5 analogs

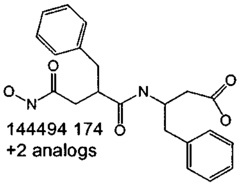
gt5 APTT

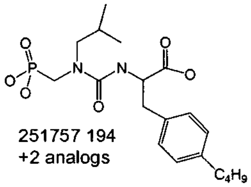

159637 49
+5 analogs


170965 54


163573 105

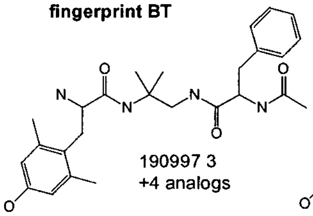

190997 106
+3 analogs

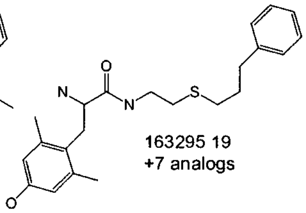

144494 174
+2 analogs

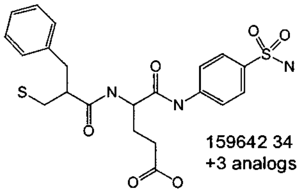


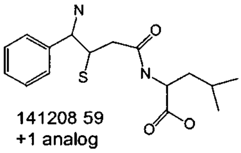
251757 194
+2 analogs

fingerprint BT


190997 3
+4 analogs

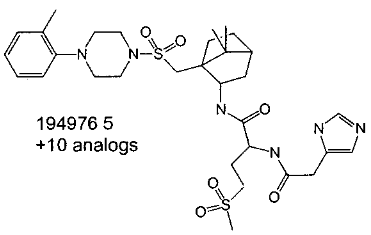

163295 19
+7 analogs

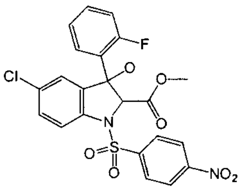

159642 34
+3 analogs


141208 59
+1 analog

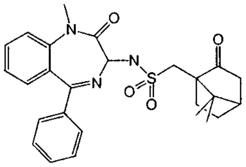
170958/oxytocin antagonists

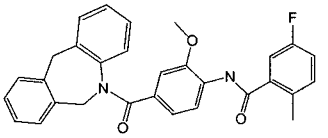
gt5 BPBT

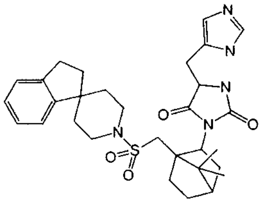

194976 5
+10 analogs



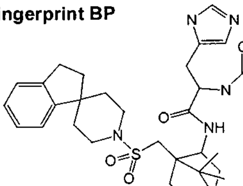
185106 23
+18 analogs


174456 54
+1 analog

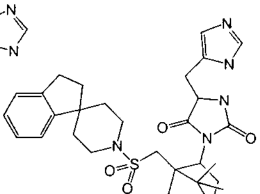

224125 135
+6 analogs


214245 164
+1 analog

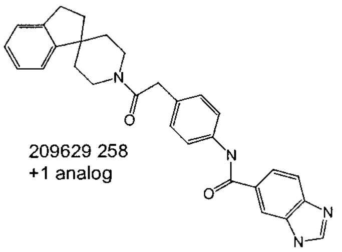
fingerprint BP



197301 7
+4 analogs



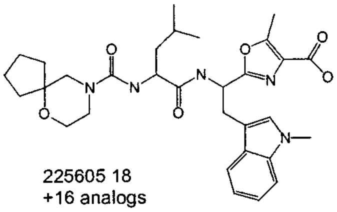
214245 49
+5 analogs



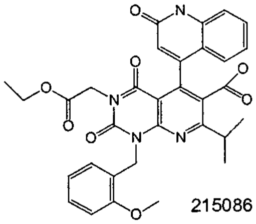
209629 258
+1 analog

174556/endothelin antagonists

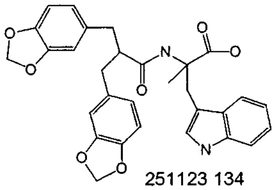
default BP



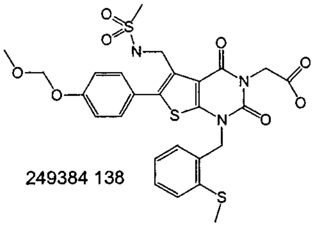
225605 18
+16 analogs



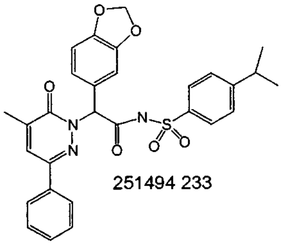
215086 128



251123 134



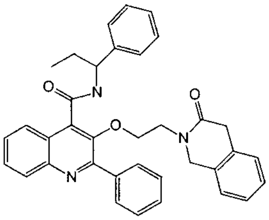
249384 138



251494 233

187236/neurokinin antagonists

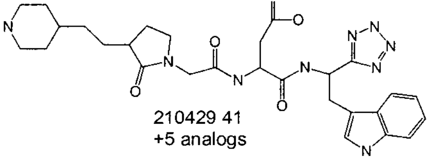
gt10 BPBT



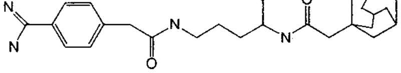
253311 27
+ 13 analogs

188541/gpIIb/IIIa antagonists

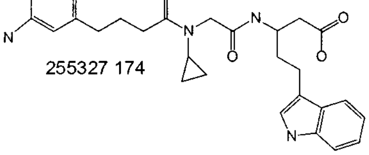
default AP



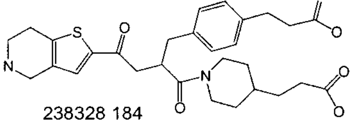
210429 41
+5 analogs



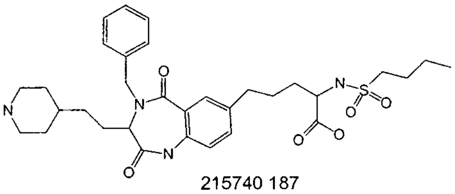
203821 136
+5 analogs



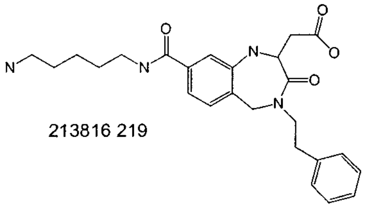
255327 174



238328 184

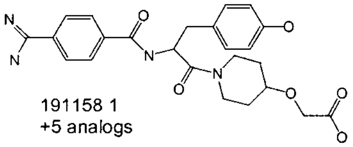


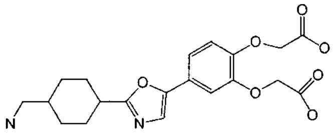
215740 187
+3 analogs

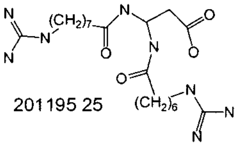


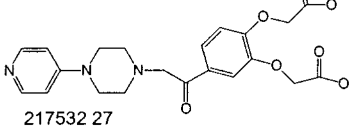
213816 219

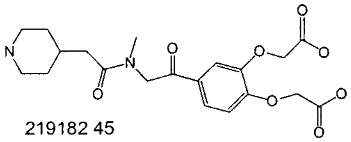
backbone BP

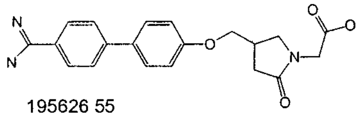

191158 1
+5 analogs


219183 18


201195 25

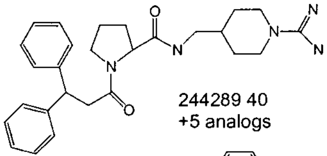

217532 27

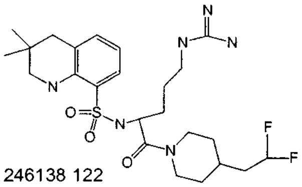

219182 45

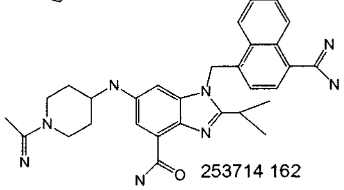

195626 55

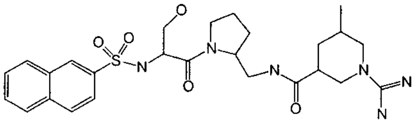
191435/thrombin inhibitors

default BPBT

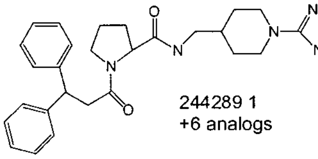

244289 40
+5 analogs

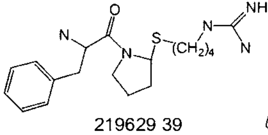

246138 122
+2 analogs

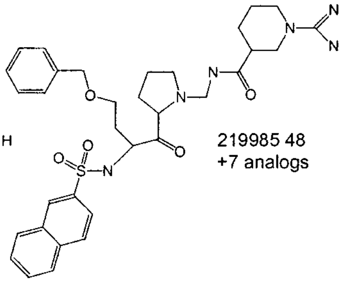

253714 162

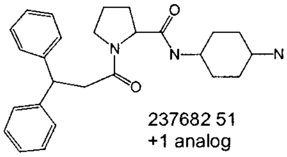

219984 216
+5 analogs

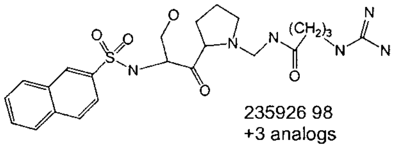
rankratio AP


244289 1
+6 analogs


219629 39

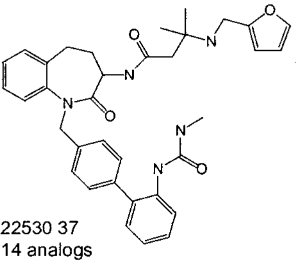

219985 48
+7 analogs


237682 51
+1 analog


235926 98
+3 analogs

195636/GH release promoting agents

fingerprint BP


222530 37
+14 analogs

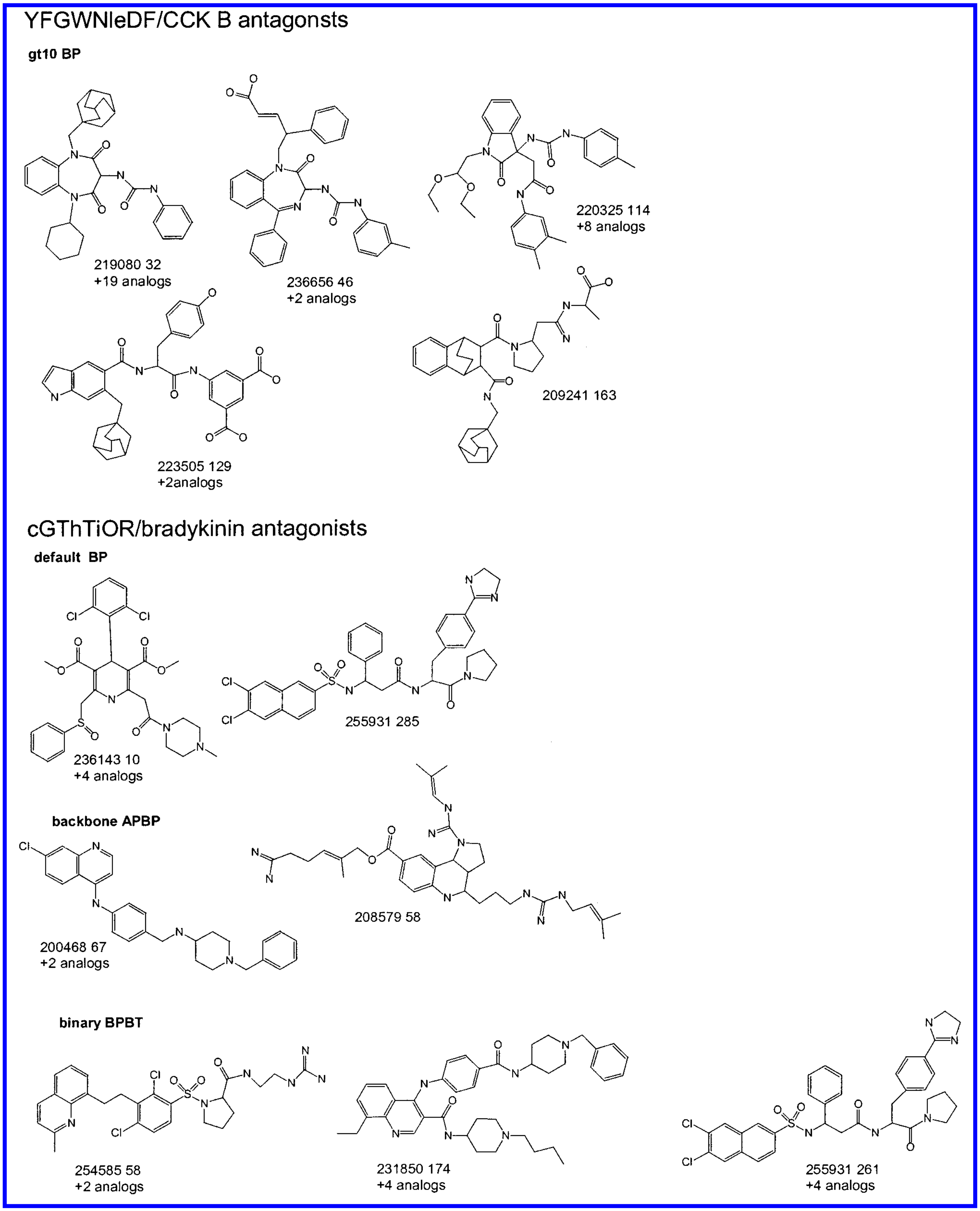


Figure 3. Selected actives from the better descriptor/protocol combinations. Each drawing is labeled by the MDDR external registry number and its rank. Also indicated is the number of analogues of that compound that have a larger rank where the rank of each analogue is ≤ 300.

of a database, one for drug-like probes and one for peptide probes, would add some maintenance overhead, and modifying the database descriptors on the fly would significantly increase the search time. Of course, this is not to say that a further improvement might not be possible by changing the database descriptors, only that it does not seem necessary to obtain acceptable results.

The most powerful protocol is to simply filter undesirable molecules from the database. In the case of MDDR, it is sufficient to eliminate peptides using a simple substructure definition. For large industrial databases, in which there may be many nonpeptides that are very nondrug-like (e.g. because they have very large molecular weights, are extremely hydrophobic, are extremely long, etc.), it may be necessary to apply other filters.

There is a limit to the length of peptides that can be handled by topological similarity searching. We obtained good results for up to heptapeptides. It is much harder to find nonpeptide actives for longer peptides (data not shown). This is not surprising for two reasons:

1. A longer probe has many more long-range AP and BP descriptors and thus would select for longer molecules in the database, unless the "gt" protocols are used.

2. A smaller fraction of chemical groups in the probe is relevant to activity.

What level of enhancement makes a similarity method worth using is, of course, a matter of judgment and depends on the throughput of biological testing to be done, but we feel that an enhancement of ∼5 is sufficient in most cases. At least one protocol seems to be able to reach that level for a majority of peptide probes. As we have seen with drug-like probes,[9] which descriptor combination to use is not obvious a priori. Here we have an extra complication that enhancements also seem sensitive to the protocol as well as to the descriptor. The only way of handling this situation is that several protocol/descriptor combinations should be run for any given probe. Fortunately, topological similarity searching is fast enough that many combinations can be run in a reasonable time. While it is hard to make a definitive recommendation, it would be reasonable to run at least three protocols, **gt10**, **fingerprint**, and **backbone**, with at least two combination descriptors, APBP and BPBT.

ACKNOWLEDGMENT

The author thanks members of the Molecular Systems Applications Group who provided in-house problems on which to test the protocols: Dr. J. Chris Culberson, Dr. Sookhe Ha, Dr. Georgia McGaughey, and Mr. Ralph Mosley. Dr. Robert Nachbar wrote **simtab**, the similarity software used in this study.

REFERENCES AND NOTES

(1) *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; John Wiley & Sons: New York, 1990.

(2) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 983−996.

(3) Miller, M. D.; Sheridan, R. P.; Kearsley, S. K. SQ: a program for rapidly producing pharmacologically relevant molecular superpositions. *J. Med. Chem*. **1999**, *42*, 1505−1514.

(4) Rohrer, S. P.; Birzin, E. T.; Mosley, R. T.; Berk, S. C.; Hutchins, S. M.; Shen, D.-M.; Xiong, Y.; Hayes, E. C.; Parma, R. M.; Foor, F.; Mitra, S. W.; Degrado, S. J.; Shu, M.; Klopp, J. M.; Cai, S.-J.; Blake, A.; Chan, W. S.; Pasternak, A.; Patchett, A. A.; Smith, R. G.; Chapman, K. T.; Schaeffer, J. M. Rapid identification of subtype-selective agonists of the somatostatin receptor through combinatorial chemistry. *Science* **1998**, *282*, 737−740.

(5) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudinere, R. F. New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged structures. *J. Med. Chem.* **1999**, *42*, 3251−3264.

(6) Jain, A. N. Morphological simliarity: a 3D molecular similarity method correlated with protein−ligand recognition. *J. Comput.-Aided Mol. Design* **2000**, *14*, 199−213.

(7) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure−activity studies: definition and application. *J. Chem. Inf. Comput. Sci*. **1985**, *25*, 64−73.

(8) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological torsions: a new molecular descriptor for SAR applications. comparison with other descriptors. *J. Chem. Inf. Comput. Sci*. **1987**, *27*, 82−85.

(9) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R.P. Chemical similarity using physiochemical property descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118−127.

(10) Bush, B. L.; Sheridan, R. P. PATTY: a programmable atom typer and language for automatic classification of atoms in molecular databases. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 756−762

(11) Hull, R. D.; Fluder, E. M.; Singh, S. B.; Nachbar, R. B.; Kearsley, S. K.; Sheridan, R. P. Chemical similarity searches using Latent Semantic Structural Indexing (LaSSI) and comparison to TOPOSIM. *J. Med. Chem*. **2001**, *44*, 1185−1191.

(12) 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., Modelling* **2000**, *18*, 343−357.

(13) MDL Drug Data Report licensed by Molecular Design Ltd., San Leandro, CA.

(14) Spear, K. L.; Brown, M. S.; Reinhard, E. J.; McMahon, E. G.; Olins, G. M.; Palomo, M. A.; Patton, D. R. Conformational restriction of angiotensin II: cyclic analogues having high potency. *J. Med. Chem*. **1990**, *33*, 1935−1940.

(15) Nikiforovich, G. V.; Kolodziej, S. A.; Nock, B.; Martinez, N. B. J.; Marshall, G. R. Conformationally readdressed CCK-B/delta-opioid peptide ligands. *Biopolymers* **1995**, *36*, 439−452.

(16) Thurieau, C.; Feletou, M.; Hennig, P.; Raimbaud, E.; Canet, E.; Fauchere, J.-L. Design and synthesis of new linear and cyclic bradykinin antagonists. *J. Med. Chem*. **1996**, *39*, 2095−2101.

(17) Daylight Chemical Information Systems, Inc, 2740 Los altos, Suite #360, Mission Viejo, CA 92691.

(18) McGregor, M. J.; Pallai, P. V. Clustering large databases of compounds: using the MDL "keys" as structural descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443−448.

(19) Heritage, T. W.; Lowis, D. R. Molecular hologram QSAR. In *Rational Drug Design: Novel Methodology and Practical Applications*; ACS Symposium Series 719; Parrill, A. L., Reddy, M. R., Eds.; American Chemical Society: 1999; pp 212−225.

(20) Sheridan, R. P. The centroid approximation for mixtures: calculating similarity and deriving structure−activity relationships. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 1456−1469.

CI0100144