Using Tversky Similarity Searches for Core Hopping: Finding the Needles in the Haystack

Stefan Senger*

GlaxoSmithKline, Medicines Research Centre, Gunnels Wood Road, Stevenage SG1 2NY, United Kingdom

Received March 9, 2009

The combination of Daylight fingerprints and the Tversky coefficient is a powerful method for performing core hopping, that is, scaffold (or lead) hopping where the main structural difference between the query and bioactive target molecule is located in the central core of the molecular structure. However, a major disadvantage of this approach is the fact that a large number of false positives (in the context of core hopping) are retrieved. The tool we have developed and which is described here can be used to postprocess the hits from Daylight Tversky similarity searches by fragmenting the molecules and subsequently annotating them in a way that assists the users in removing false positives and enables them to better focus on molecules of interest. To validate our approach, we have selected four biological targets for which scaffold hopping examples have been reported. We present results from searches in databases containing published activity data and the subsequent analysis of the hits aimed at establishing the potential of our approach.

INTRODUCTION

Scaffold hopping[1,2] is an important strategy in ligand-based hit generation. The term was coined by Schneider et al. and describes an effort which is aimed at "identification of isofunctional molecular structures with significantly different molecular backbones".[3] In the simplest case, it is sufficient to have just one molecule that can be used as a starting point (template, reference molecule) for the scaffold hopping exercise.

Most often scaffold hopping (which is also sometimes referred to as lead hopping)[4] is focused on identifying molecules that have a significantly different structural makeup irrespective of the fact whether or not the structural difference is localized. However, there are cases when it is desirable to only change defined structural regions within a molecule. An example might be a situation where the terminal groups of a molecule are perceived as being crucial for binding to the receptor and modifications of the molecular structure are focused on the core region, that is, the part of the molecule that joins the terminal groups. This specific scaffold hopping (or lead hopping) scenario is referred to as core hopping.[5] To illustrate this, two Bradykinin 1 antagonists, **1** and **2**, are shown in Figure 1.

Apart from the presence of two additional fluorine atoms, the terminal groups of **1** and **2** are the same and the structural difference is localized in the core of the molecular structure. Despite the significantly different cores both molecules show nanomolar activities against human Bradykinin 1.[6,7]

It is obvious from the above that it will be relatively straightforward to implement a core hopping algorithm by using substructure searching, for example, as long as only molecules that contain an exact structural match for the terminal groups are considered as hits when searching databases of interest. However, it may also seem desirable to retrieve molecule like **3**,[8] which will not be found when an exact substructure search is performed. A possible strategy to also capture hits like **3** would be to perform searches using molecular patterns instead of exact substructures, for example, Daylight[9] SMARTS. A clear disadvantage of this approach is that results will strongly depend on the definition of the molecular pattern used for the search and might therefore be strongly biased. A further disadvantage is the lack of a suitable metrics to indicate how closely related the terminal fragments in the query molecule and the hit are. This observation has prompted us to investigate the use of a combination of a fingerprint-based molecular descriptor and similarity index. We have chosen a combination of the widely used Daylight 2D-fingerprints[9,10] and the Tversky similarity index.[11,12] The Tversky index is an asymmetric index, which contains two user-defined parameters (α and β).[13] If both of these parameters are set to 1 it becomes the popular Tanimoto[12] index. On the other hand, if α is set to 1 and β is set to 0, the Tversky index is a measure for substructural similarity, where a Tversky index of 1.0 indicates that a given structural moiety is a substructure of the molecule with which it is being compared.

For example, the Tversky index (with $\alpha = 1$ and $\beta = 0$) of **2** when compared with 3,3,3-trifluoropropionamide (Frag1, green) and 4′-aminomethyl-[1,1′-biphenyl]-2-carboxylic acid methyl ester (Frag2, blue) is 1.0 since both structural moieties are contained as substructures in **2** (cf. Figure 2). This example nicely illustrates the potential of the Tversky index to be used for core hopping. The Daylight Tanimoto similarity value for **1** and **2** included in Figure 2 shows that it is extremely unlikely that **2** could be found by performing a Daylight Tanimoto similarity search. The impact of the choice of similarity measure on compound ranking, for example, has recently been discussed by Rupp et al.[14] (with an emphasis on high-dimensional chemical descriptor spaces).

For completeness, it should be pointed out that a successful replacement of 2,3-diaminopyridine by cyclopropylamino

* To whom correspondence should be addressed. E-mail: stefan.x.senger@gsk.com.

10.1021/ci900092y CCC: $40.75 © 2009 American Chemical Society
Published on Web 05/19/2009

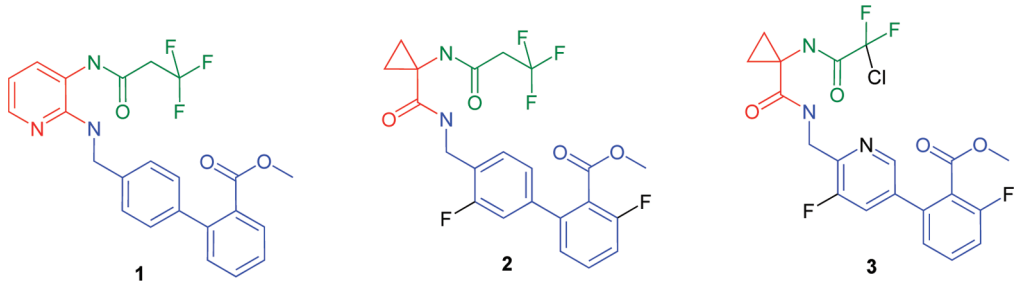


Figure 1. Structurally related Bradykinin 1 antagonists.

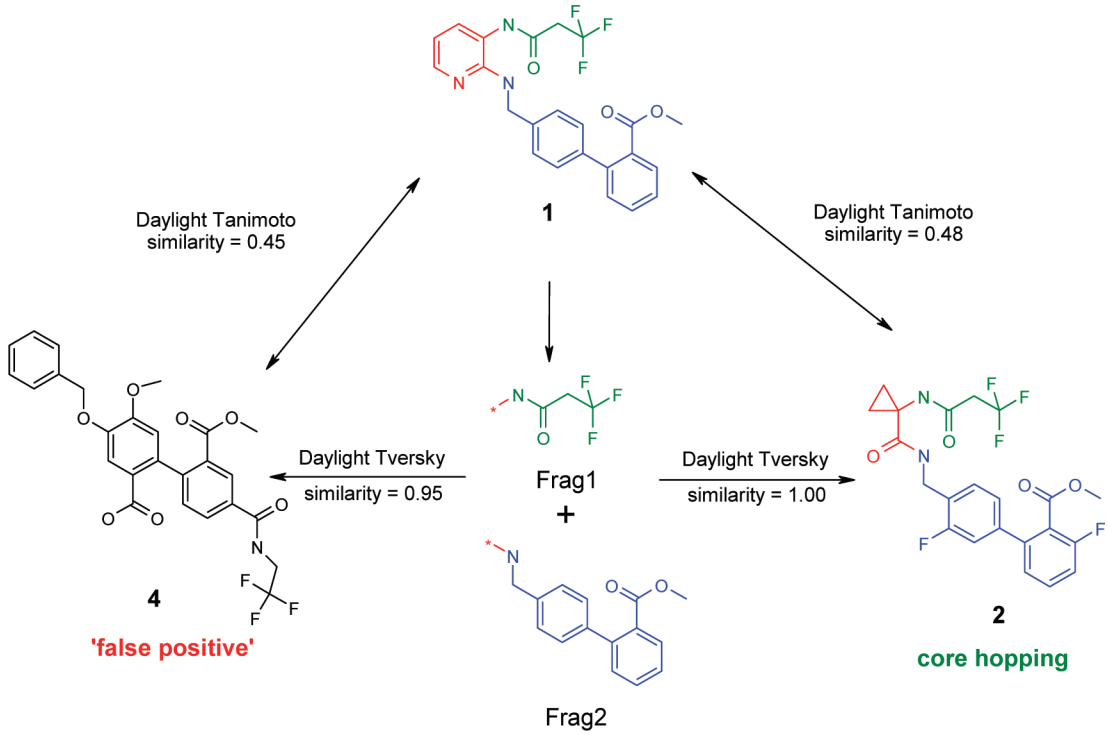


Figure 2. Examples for Daylight Tanimoto and Tversky similarity searches.

acid amide has not only be reported for Bradykinin but also for the serine protease Factor Xa.[15]

When looking through hits from Daylight Tversky similarity searches aimed at core hopping it quickly becomes obvious that one is faced with a major problem since a significant number of undesirable molecules are found among the hits. What we mean by undesirable is illustrated by **4**.[16] This molecule has a high Tversky similarity of 0.95 when compared with the two reference fragments (see Figure 2) but unlike **2** it does not contain the moieties that are similar to the two reference fragments as separate entities. In the context of what we are trying to achieve (i.e., core hopping), this molecule would have to be considered as a false positive. To enhance the potential of this approach, it would either have to be modified in such a way that the occurrence of a large number of false positives is avoided or a way to postprocess the hit list from a Daylight Tversky search would have to be implemented so that it becomes straightforward to separate the false positives from the molecules of interest. A strategy that could be used to avoid the occurrence of false positives was reported by Wagener and Lommerse.[17] They first fragment all molecules in the search space and create a database (where fragments are labeled as, for example, core or R-group) and then perform similarity searches in this database. Obviously, the choice of the fragmentation rules that are used has a significant influence on the results obtained from the subsequent searches. While this approach will make sure that hits contain the correct connectivity, it requires a fragment database that can be searched. If the search space is the compound file of a large pharmaceutical company, this can represent a major logistical challenge. This perceived disadvantage has motivated us to explore ways of postprocessing the hits from Daylight Tversky searches aimed at developing a tool that provides a way to guide the user toward the hits that are of greatest interest in regards to core hopping.

The approach we have chosen, and which will be described here, is to apply an algorithm to the hits that selects (single) bonds and subsequently breaks them resulting in a number of molecular fragments. A condition for a valid fragmentation is that all terminal fragments are joined via a common central fragment, which we refer to as core fragment. The (single) bonds that are broken by the algorithm are selected in such a way that the resulting terminal fragments have a maximum Tversky similarity when compared to the query fragments (i.e., the terminal fragments derived from a known active

and used for the initial Tversky search, cf. Figure 2). The perceived advantage of this approach is that it is also similarity based and hence only requires the query fragments from the initial Tversky search as input. While it would be simpler to apply a substructure-based fragmentation (e.g., through specific SMARTS[9] matching) this would introduce a further bias through the way the substructures for the fragmentation are defined and would thereby significantly hamper the power of this approach. After the fragmentation algorithm has been applied to the hits, a set of simple properties is calculated for the fragments that have been obtained. The aim is to be able to use this annotated hit list to select molecules of interest by interactive visualization, similarly to what is used to analyze the output from high-throughput screening campaigns.[18] To integrate the different steps of this approach, we have developed a tool that we have called CORUS.[19] To critically evaluate the potential of this tool, we have used it to perform searches in the Aureus Pharma AurSCOPE databases,[20] using published core hopping examples. The results of our studies are reported here.

EXPERIMENTAL METHODS

All components of CORUS have been programmed using Perl in conjunction with the following modules: (a) DayPerl[9] to interact with Daylight version 4.92, (b) DBI to query the AurSCOPE[20] or our internal Oracle databases, and (c) LWP and XML to interact with internal web services (e.g., to calculate simple properties). For more details about the individual modules, see http://cpan.search.org.

To illustrate the methodology we will use the example introduced in Figure 2.

Input. For each CORUS run, there are three types of parameters that have to be specified by the user. These are (a) the names of the databases that will be searched (which for this study will always be the collection of AurSCOPE[20] databases available to us, that is, AurSCOPE ADME/DDI, AurSCOPE GPCR, AurSCOPE hERG, AurSCOPE Ion Channels, and AurSCOPE Kinase), (b) the SMILES strings for the fragments that will be used as queries for the initial Daylight Tversky search (with the attachment points being labeled by asterisk characters; see, for example, Frag1 and Frag2 in Figure 2), and (c) the threshold for the Tversky index that is used for the initial database search. Optionally, a user can define a maximum number of hits that will be retrieved from a database and an upper limit for the heavy atom count that will be used to filter out molecules that are considered to be too large.

It is also possible to define a reference molecule (e.g., the molecule from which the query fragments used for the initial Daylight Tversky search have been derived). The structure of this molecule is not used for the initial Daylight Tversky search but will also be fragmented and the resulting core fragment (i.e., the fragment that joins the terminal fragments, referred to here as reference core) allows for Tanimoto similarities between the core fragments derived from a similarity search hits and the reference core to be calculated. Having these values available in the output offers an easy way to filter out hits containing cores that are very similar to the core in the reference structure (and therefore most probably not of interest). In case of the example introduced in Figure 2, the two substructures labeled as Frag1 and Frag2 are used as query fragments for the search. The known active **1** (from which the two query fragments have been derived) could be used as a reference molecule. If this is the case, the pyridine-containing substructure of **1** colored in red would become the reference core.

Database Search. To construct the query SMILES strings, the query fragment SMILES are concatenated together with a dot character into one query string which is subsequently used to perform the Tversky search in the databases. From the search, for each hit that is retrieved the identifier (which in this case is the Aureus Molecule identifier), the SMILES string, and the value of the Tversky similarity index are captured. If a threshold for the heavy atom count was specified, molecules that do not pass this filter are not further processed.

Fragmentation. The objective of the fragmentation algorithm is to fragment the hits in such a way that the terminal fragments that are obtained by breaking bonds are as similar as possible to the query fragments while also obtaining an additional (core) fragment that joins all terminal fragments.

To illustrate this, we refer to Figure 2 again. If **2** was a hit obtained by a Daylight Tversky similarity search using substructures Frag1 and Frag2 as query fragments, the desired outcome of the application of the fragmentation algorithm would be to obtain the two substructures colored in blue and green as terminal fragments and the substructure colored in red as core fragment.

In the approach we have chosen, an array of mutant structures is generated for each molecule that is given to the fragmentation algorithm as input. The set of mutated molecules is generated by individually replacing all heavy atoms of the input structure by a dummy atom one at a time. In our case, we have chosen arsenic as dummy atom.[21] Three examples for mutant structures generated from **1** are shown in Figure 3.

In the next step, the asterisk characters (which mark the attachment points) in the query fragments are mutated in the same way so that subsequently the Tversky similarity between the two asterisk-mutated query fragments and every member of the array containing the mutated hits can be calculated. (cf., Figure 3. Be aware that only 3 mutants out of 31 are shown.) The aim of doing this is to determine which of the atoms in the hit molecule best corresponds to the respective attachment points in the query fragments. This is simply done by selecting the mutant with the largest Tversky similarity index for each query fragment (e.g., mutant **B** for Frag1′ and **A** for Frag2′). It is apparent that the example we are using here for illustration purposes is trivial inasmuch as both query fragments are substructures of the molecule we use as an example for a Daylight Tversky hit (i.e., molecule **1**) and the largest Tversky similarity value is therefore 1.0. Very encouragingly, fragmentations we performed on a variety of molecules (with a range of query fragments) have shown that this approach also works in a reliable fashion if the query fragments are not exact substructures of the respective similarity search hit. These finding have encouraged us to implement the approach in our tool CORUS.

Not surprisingly, the more dissimilar the query fragments are to the corresponding terminal fragments in the hits, the more challenging it becomes for the fragmentation algorithm to perform what is perceived as the correct fragmentation.

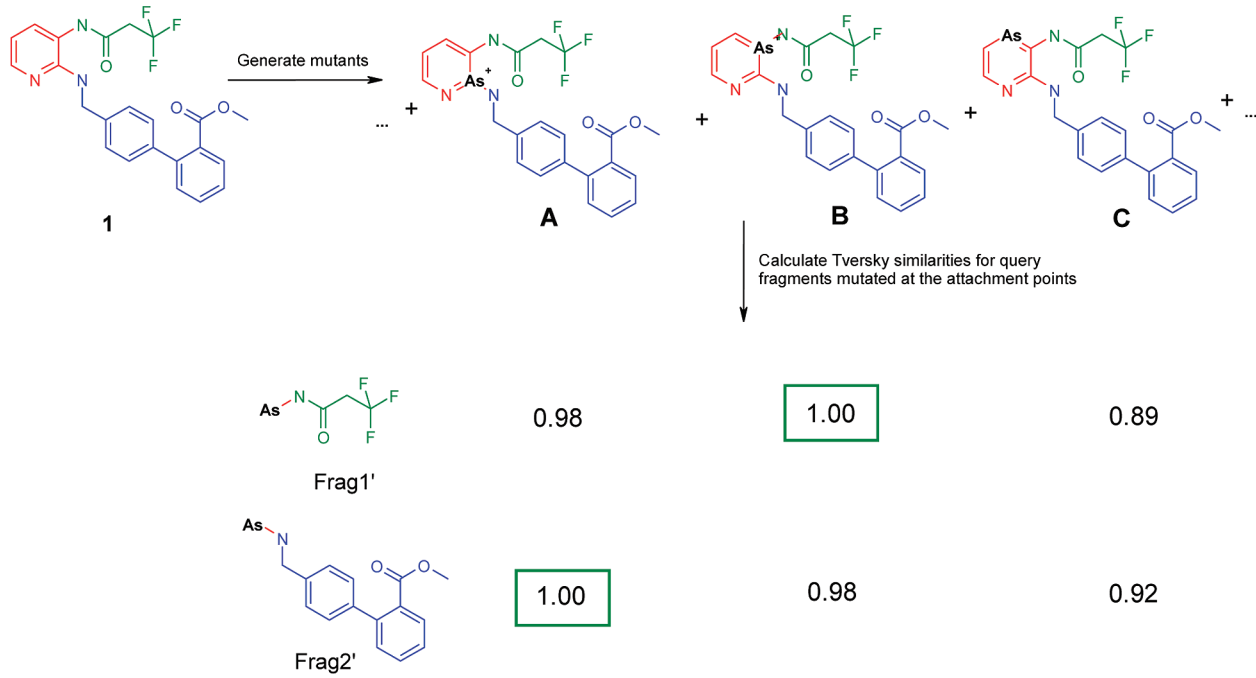


Figure 3. Illustration of the initial steps of the fragmentation algorithm used for CORUS. Only 3 (out of the 31) members of the mutant array are shown. The mutants that are further investigated in the subsequent steps of the algorithm are those with the largest Tversky similarity values (when compared with the query fragments that have been mutated at the attachment points).

We found that the failure rate is reduced, if not only mutants with the highest Tversky similarity are progressed into the next step of the algorithm but also additional high ranking matches are taken forward. Therefore, we have introduced two parameters[22] that can be used to modify this behavior. Our finding is similar to that of Bayada et al.[23] in regards to their algorithm for finding multiple common subgraphs.

It should be mentioned that the maximum path length that is defined when the Daylight fingerprints are calculated has an effect on the way the fragmentation algorithm behaves. In the fragmentation algorithm we have set the maximum path length to a value of 5 (instead of 7, which we otherwise use) when fingerprints for mutants are calculated. The reason for reducing the value for the maximum path length from the default 7 to 5 was prompted by the observation that the fragmentation algorithm performs more reliable in (what we perceived as) difficult cases when the smaller value is used.

In the subsequent step, the mutants previously selected are examined further (e.g., **A** and **B** in Figures 3 and 4). First, all atoms that share a single bond (that is not part of a ring) with the mutated atom are identified. These atoms are marked with a red circle in Figure 4. Then the corresponding mutants (i.e., **A′** and **B′** in Figure 4) are selected from the mutant array that has already been generated.

After the query fragments have been mutated at the atoms that are next to the attachment points (cf., Figure 4), Tversky similarities for the mutated query fragments and the selected mutants (e.g., **A′** and **B′**) are calculated and the mutants with the largest Tversky indices are selected (if there is more than one mutant). The latter is exactly the same as what is illustrated in Figure 3 with the exception of the position at which the query fragments are mutated (see Figures 3 and 4 and compare the mutated query fragments that are shown). By following the algorithm described here, it becomes possible to identify the bonds that most likely result in the desired fragments to be formed when they are broken. Again, it has to be highlighted that the example used here for illustration purposes is a trivial case since there is only one mutant per query fragment (i.e., **A′** and **B′**, respectively), and therefore, no selection based on Tversky similarity is required.

In some cases the algorithm might identify more than one bond per query fragment. If this is the case, all feasible combinations of bonds are generated and the actual fragmentations are performed. Furthermore, the Tversky similarities relative to the corresponding query fragments are calculated for the terminal fragments obtained by these fragmentations. The winning fragmentation is the one where the sum of the Tversky similarity values is the largest. For completeness, it should be mentioned that even if the algorithm discussed here has to perform more than one fragmentation per molecule it is still much more efficient than using the brute force approach, i.e. performing all possible fragmentations and subsequently selecting the "winning fragmentation" by calculating the Daylight Tversky similarities.

Apart from the SMILES for the fragments obtained by the fragmentation and the corresponding Tversky similarities, the heavy atom counts for the fragments are also calculated. Furthermore, an error flag is returned, indicating whether or not it was possible to fragment the hit molecule into terminal fragments and a corresponding core fragment.

In the implementation of this algorithm in the version of CORUS that was used for this study only two query fragments can be provided as input and the fragmentation will consequently result in two terminal fragments and a core fragment (if possible).

Further Annotation. Although we are not necessarily looking at alternative cores that are similar to a known core

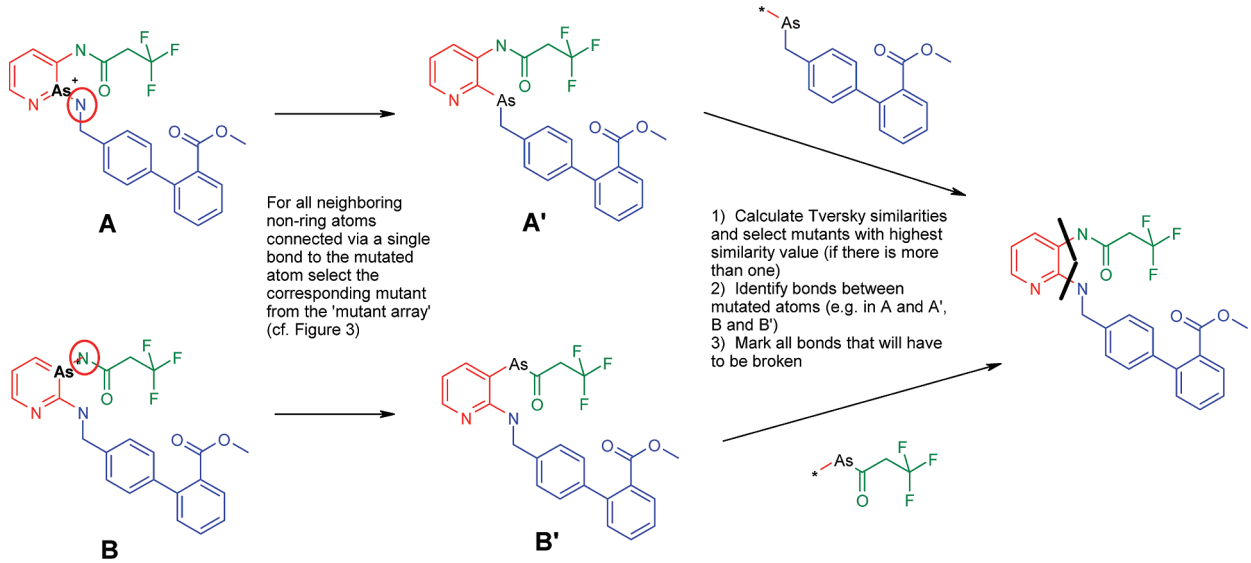


Figure 4. Illustration of the final steps of the fragmentation algorithm used for CORUS. For each mutant selected in the previous step all atoms that are not in a ring and are connected to the mutated position via a single bond are identified. For each of these atoms the corresponding mutant from the mutant array (cf., Figure 3) is selected. This particular example is a trivial case since there is only one such atom in **A** as well as in **B** (and the corresponding mutants are shown as **A′** and **B′**). If this is not the case a similar approach to what is shown in Figure 3 is used, that is, the Tversky similarities between the newly selected mutants and the query fragments (mutated next to the attachment point) are calculated and the mutant with the largest similarity value is selected. By using the information from the mutants (e.g., **A**, **A′**, **B**, **B′**) it becomes possible to identify the bonds that have to be broken to achieve the desired fragmentation.

(e.g., the core in the optional reference molecule) in Daylight Tanimoto or Tversky fingerprint space, we might still have some particular requirements for cores we consider to be of interest. One consideration will be the distance between the terminal fragments which are linked through the core. For this study, we decided to use a simple descriptor for the distance, which is the number of atoms on the shortest path between two terminal fragments. To calculate this, we use a modified version of the Perl script minpath.pl that is one of the DayPerl examples that can be obtained from the Daylight[9] Web site. We also calculate a simple path descriptor which captures for every bond on the path if it is a member of a ring, if it is aromatic, and if it is rotatable. This descriptor is used to cluster the hits of the Tversky search, which is hoped to be helpful when visually inspecting hits as part of the analysis.

Additionally, the following simple descriptors are calculated for the cores: (a) hydrogen-bond acceptor and donor counts, (b) number of aliphatic and aromatic rings, and (c) number of rotatable bonds (in the core as well as on the shortest path). The values are obtained by calling a web service.

The last step in CORUS is optional and is only performed if a search has been performed in an internal database. If this is the case and the user has specified a biological assay, the corresponding assay data is retrieved (by querying the Oracle database containing the screening data) and added to the CORUS output. The final output is written into a comma-separated values (csv) file which the user can then analyze.

RESULTS AND DISCUSSION

To establish if CORUS is performing as expected in regards to the purpose it has been developed for (i.e., primarily core hopping) and to evaluate its potential, we have used examples from the literature where molecules from the relevant biological target are captured in one or more of the AurSCOPE[20] databases.

Bradykinin. For the initial test, we chose the Bradykinin example used above. The fragments derived from **1** (cf., Figure 2) together with a (generous) Tversky threshold of 0.75 resulted in approximately 11 500 hits when we searched in the AurSCOPE databases available to us (using a maximum heavy-atom count of 50 for the hits and a maximum number of hits per database of 10 000). Figure 5, nicely illustrates the issue that prompted the development of CORUS, that is, the fact that the Tversky index for a hit from the initial similarity search is high does not necessarily imply that it contains fragments that are similar to the query fragments (and connected by a core fragment). For example, for hits from this search where the Tversky index from the initial similarity search is ≥ 0.9 only approximately 60% have a mean Tversky index of ≥ 0.7 for the terminal fragments (obtained by applying the fragmentation algorithm). For similarity search hits with a Tversky index of ≥ 0.85, this is already reduced to about only 25% and is even as low as 10% when the hits with a Tversky index of ≥ 0.8 are considered.

It has to be stressed that this analysis can not be entirely accurate since there are a number of situations when CORUS might/will fail, distorting the numbers quoted above. The two most likely situations for a failure are that either a fragment is connected to the core with more than one bond (i.e., it is part of a ring) or the fragments are connected together directly (lacking a core). It is arguable that the latter case is a failure, since the main objective here is core hopping so that compounds lacking a core are of minor interest. However, at least based on the experiences we have gained by performing these studies such cases are uncommon and do not significantly impact on the final outcome.

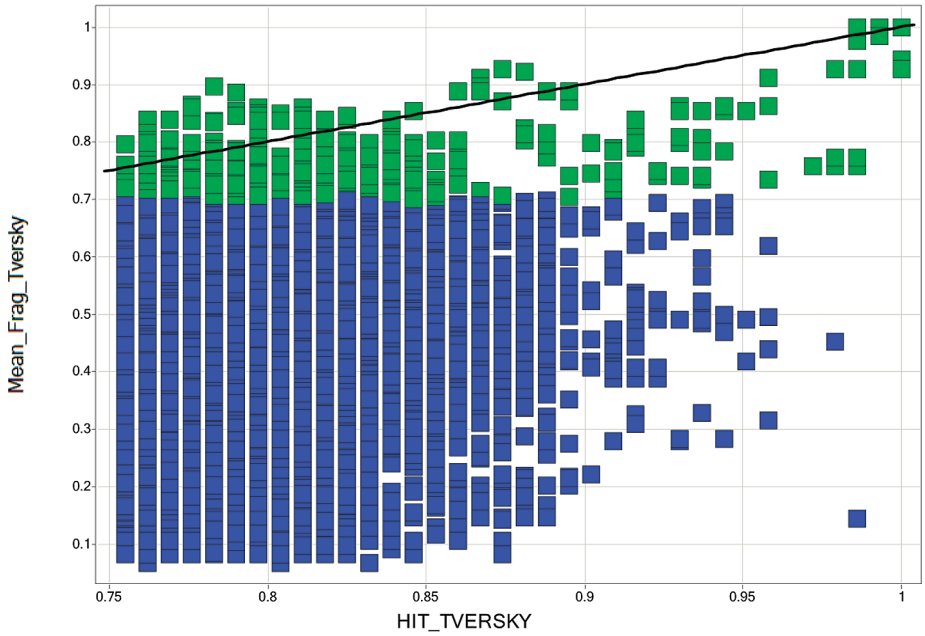
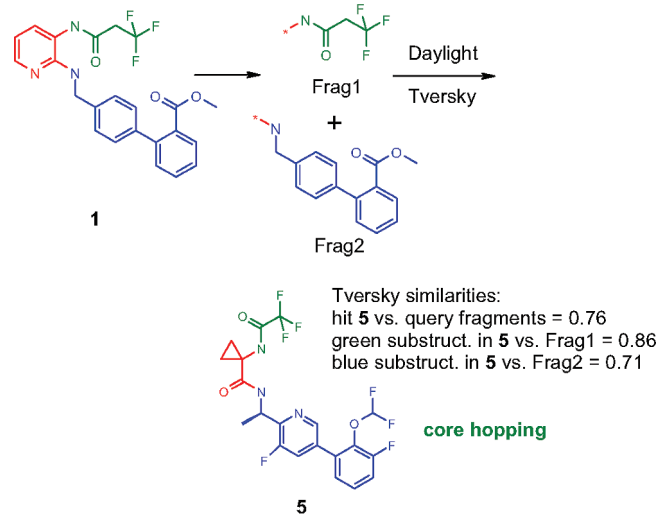


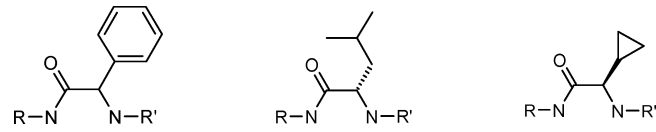
Figure 5. Mean Tversky index for the (terminal) fragments obtained by fragmentation of the hits (when compared to the matching query fragments) plotted against the Tversky index from the initial similarity search. Hits with a mean Tversky index of ≥0.7 for the hit fragments are colored green.

As discussed earlier, we try to annotate the CORUS output in such a way that a user will hopefully be able to easily focus on the hits of interest by using an interactive visualization tool (e.g., Spotfire[24]). To trial this approach, we have used the information contained in the CORUS output to filter the hits from the initial similarity search. It has to be stressed that the criteria we have used for the filtering are arbitrary to a certain extent since they can be freely chosen by the individual user. However, we see this as a crucial component of this approach (i.e., giving the user flexibility) and nothing is lost by choosing a particular set of filter criteria since they can be modified and adjusted in subsequent cycles of analysis (if this is what the user decides to do). For the current example, choosing a threshold of 0.7 for the Tversky similarity values for the two fragments already reduces the number of hits from 11 500 to only approximately 470. After application of a size filter for the fragments (i.e., a heavy atom count range of 5−15 for fragment 1 and 12−30 for fragment 2, both based around the corresponding value in the query fragments) and the core (i.e., the core should not have less than 3 and more than 15 heavy atoms), as well as application of a cutoff to the length of the path that connects the two fragments (i.e., the path should contain at least 2 but not more than 4 atoms), 266 hits remain. This number was deemed to be small enough to proceed with the analysis by visual inspection (e.g., by stepping through the path clusters, see above). To check the reliability of CORUS in regards to finding obvious hits, we looked at the 69 hits (among the 11 500) that have a Tanimoto similarity of ≥0.85 when compared with the reference molecule **1** from which the query fragments were derived. Encouragingly, 63 of them have passed the above filters and are part of the set of 266 hits. The remaining six were examined visually, and there are good reasons why they failed the filters. We have taken this as an indication that the splitting algorithm implemented in CORUS produces the expected results.

Substructure searching reveals that 172 of the 266 remaining hits contain the same core as the reference molecule **1**. Forty-two hits contain the same core as **2** (with **2** being one them). One of these hits is **5**.[8] It is encouraging to see that by finding this molecule we could have identified the alternative core reported in the literature[15] even though the terminal fragments in this molecule are already distinctly dissimilar in comparison to the query fragments and it would have been difficult to find this molecule through a substructure-based search.



Three examples of alternative cores[25] that have also been found are shown below.



Cyclooxygenase-2 (COX-2). Our second core hopping example is from the area of COX-2 inhibitors. In 1995 Gauthier et al.[26] reported their efforts to replace the thiophene core in DuP 697 (in order to address oral absorption issues).

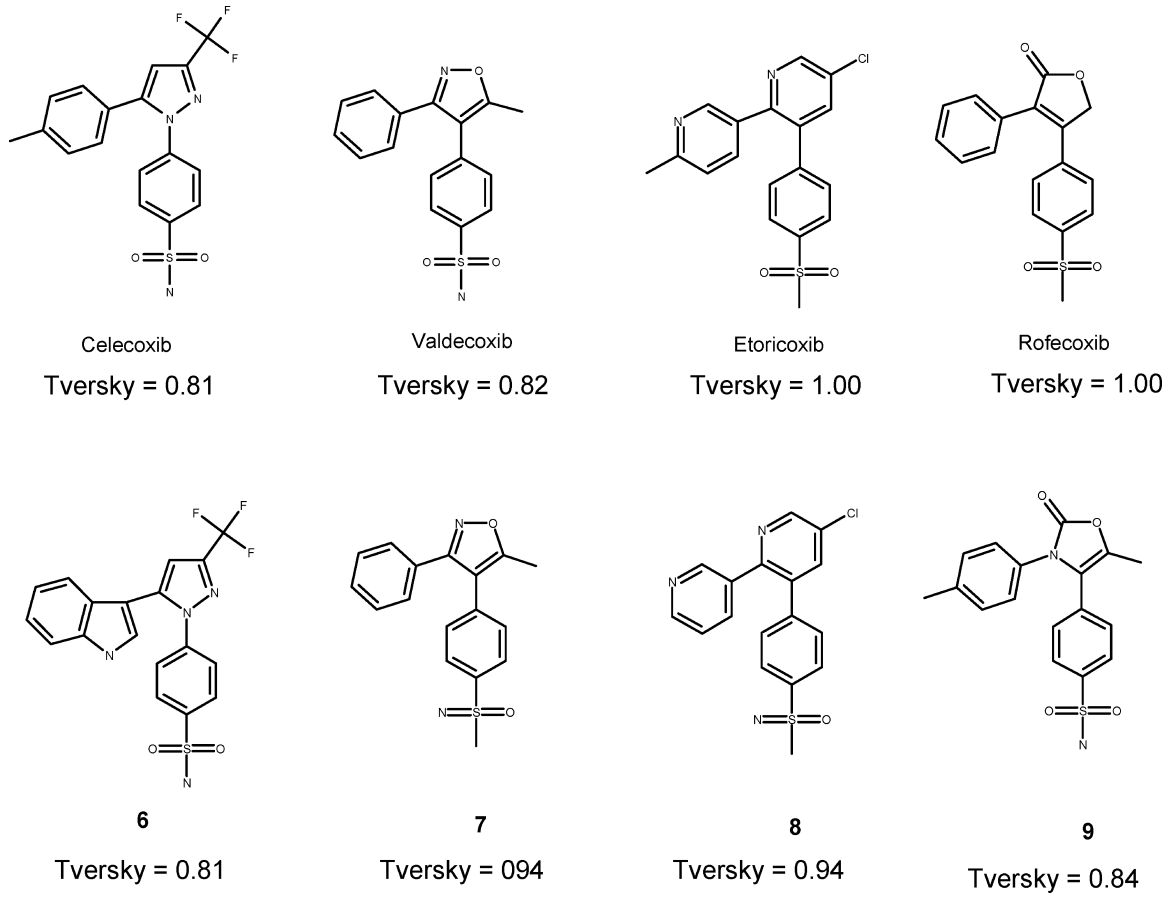
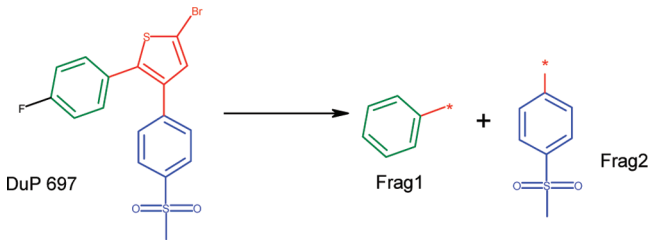


Figure 6. Subset of hits with reported COX-2 inhibitory activity. All molecules shown have a Tanimoto index of ≤0.35 (when compared with DuP 697).

They retained the methylsulfonylphenyl moiety but decided to use a 4-fluorophenyl as well as a phenyl group as second terminal fragment (linked by alternative cores). Following this example, we chose methylsulfonylphenyl and phenyl as the two query fragments.



A search in the AurSCOPE databases using a Tversky threshold of 0.80 resulted in approximately 12 700 hits (when a maximum heavy-atom count of 35 for the hits and a maximum number of hits per database of 10 000 were used). Application of a Tversky similarity threshold of ≥0.7 for the two fragments generated by CORUS reduced the number of hits to roughly 4000. DuP 697, which was used as reference molecule, has the following properties: (a) the heavy atom counts (HAC) are core = 6, frag1 = 7, frag2 = 10, (b) the length of the core atom path is 2, and (c) the core contains one ring. These properties were used to guide the choice of filters for the CORUS hits. After application the following constraints, only 85 hits remained: (a) HAC-core = 5−10, frag1 = 5−10, frag2 = 5−15, (b) length of core atom path = 2, and (c) the core must contain at least one ring. Since currently there is no AurSCOPE database that captures cyclooxygenase activity, we could not easily determine how many of the 85 hits are reported to have COX-2 activity. However, literature searches revealed that at least 27 of the 85 hits appear to have some level of COX-2 inhibitory activity.

A subset of molecules reported to have COX-2 inhibitory activity are shown in Figure 6. Rofecoxib[27] is an example for a trivial hit because it contains the query fragments and it would therefore have been straightforward to find this molecule through a substructure search. Etoricoxib[28] is a hit where only fragment 2 is an exact match whereas fragment 1 contains a pyridine instead of a phenyl ring. The reverse is true for Celecoxib,[29] Valdecoxib,[30] **7**,[31] and **9**.[32] These hits all contain modified groups matching fragment 2. The most interesting hits in the context of what we are trying to achieve here are the ones with modifications in both fragments. It is encouraging to see that CORUS was able to find **8**,[31] as well as **6**.[33] The latter only has a Tversky similarity of 0.81 in the initial similarity search, and it would have been very tedious to find this molecule among the almost 13 000 hits.

Cannabinoid Receptor 1 (CB_1). Rimonabant[34] is considered to be the first CB_1/CB_2 subtype-selective CB_1 antagonist. It first appeared in a publication in 1994. Since then, Rimonabant has been used as a template for a variety of scaffold hopping exercises, whereby a particular focus has been placed on identifying suitable replacements for the central pyrazole moiety.[35] Since in its current implementation

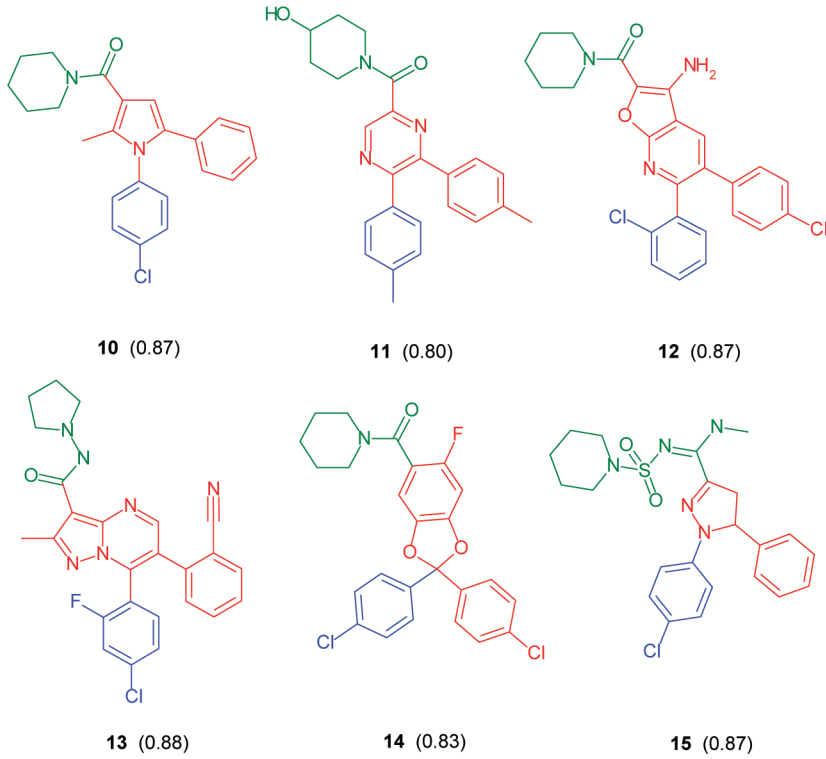
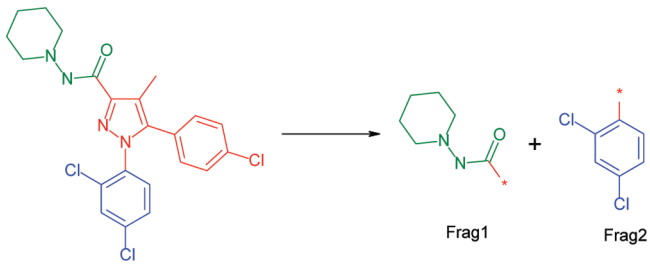


Figure 7. Subset of hits with reported CB_1 activity. The Tversky indices for the initial similarity search are shown in brackets. All molecules shown have a Tanimoto index of ≤0.55 (when compared with Rimonabant).

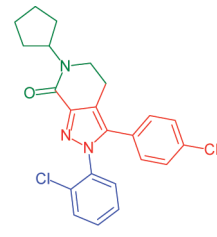
CORUS's splitting algorithm requires two query fragments, we had to make a decision, which two out of the three groups that decorate the pyrazole ring in Rimonabant to use. For the search we performed, we decided to use the *N*-1-piperidinylamide moiety (fragment 1) together with the 2,4-dichlorophenyl group (fragment 2).



A search in the AurSCOPE databases using a Tversky threshold of 0.80 resulted in approximately 25 300 hits (when a maximum heavy-atom count of 45 for the hits and a maximum number of hits per database of 30 000 were used). To be able to find hits of interest that contain groups that match fragment 2, other than chloro-substituted phenyl groups, we chose a lower than usual Tversky threshold of ≥0.6 (for both fragments). Approximately 11 700 hits passed this filter. This number was further reduced to roughly 3350 after application of the following additional filters: (a) no more than 15 heavy atoms in fragment 1, (b) no more than 10 heavy atoms in fragment 2, and (c) at least one aromatic ring in the core fragments. Encouragingly, approximately 1200 out of the 3350 molecules have CB_1 data associated with them in AurSCOPE GPCR database. Since the main aim of this study is to evaluate the potential of CORUS to identify hits that most likely would not have been found by performing substructure searching, we have focused on hits that do not contain an exact match for both query fragments.

A diverse selection of hits with alternative cores is shown in Figure 7. **10**[36] is an example for a hit where the pyrazole ring is replaced by a different five-membered aromatic ring. Similarly, **11**[37] exemplifies the class of CB_1 antagonists with a six-membered aromatic ring in the center of the molecule. A further step away from the pyrazole ring is the use of 5,6-fused aromatic rings as, for example, in **12**[38] and **13**.[39] Another example with a fused ring system is **14**.[40] **15**[41] is the only hit shown in Figure 7 where fragment 1 is distinctly different from the corresponding query fragment. The five other examples contain groups that are closely related to the *N*-1-piperidinylamide moiety. The variance observed for fragment 2 is even smaller, which is in line with the tight SAR reported for this group.[35] Generally speaking, our observations are in line with the fact that most scaffold hopping efforts starting from the Rimonabant scaffold have been core hopping exercises.[35]

It was mentioned that in its current implementation the fragmentation algorithm in CORUS is not able to cope with situations where a matching fragment is part of a ring. Among the CB_1 hits is a molecule, **16**,[42] that nicely illustrates



this case. The moiety that matches query fragment 1 (shown in green) in **16** is connected to the core via two single bonds

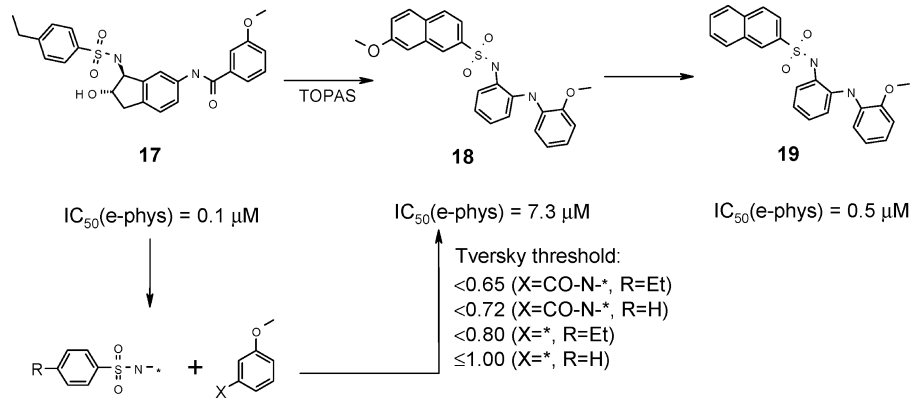


Figure 8. Scaffold hopping example reported by Schneider et al.[43] Activities are related to human Kv1.5. Tversky threshold are given for different query fragment combinations. Only when the given Tversky threshold is used for a similarity search with the corresponding query fragments **18** could be found among the hits (if it is contained in the database that is searched).

that are part of a six-membered ring. Since currently ring bonds are not considered when splitting a molecule into fragments, the fragmentation algorithm fails (i.e., splits in a suboptimal way). Future developments will seek to address this deficiency.

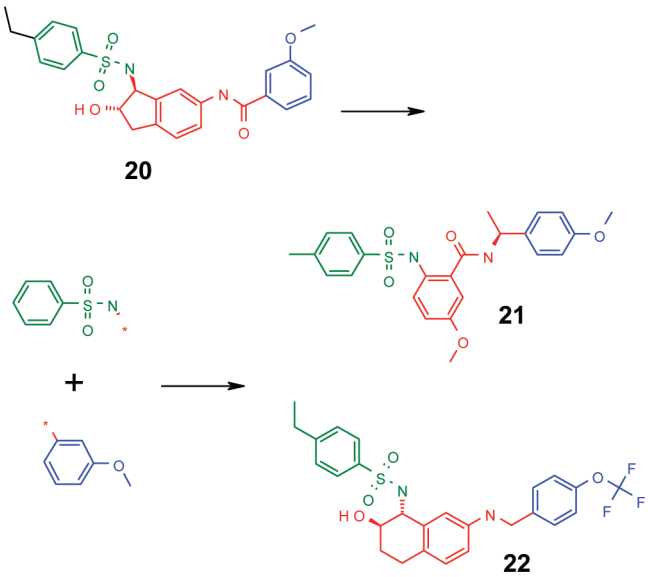
Voltage-Gated Potassium Channel Kv1.5. Schneider et al. reported a scaffold hopping example where they used a pharmacophore-guided evolutionary searching approach (called TOPAS, which stands for topology-assigning system).[43] As a starting point they selected the human Kv1.5 inhibitor **17**. This molecule is reported to have an IC_{50} of 0.1 μM in a whole-cell patch clamp electrophysiology assay and is the most potent compound for which activity data is reported (in this assay) in ref.[44]

Using TOPAS, Schneider et al.[43] identified **18** and improved the level of activity by a minor structural modification (resulting in **19**) to roughly the same level as observed for the query molecule. We decided to use this example to explore the influence the choice of query fragments has on the outcome of the Tversky similarity search. First, we chose 2-methoxyphenylamide and 4-ethylphenylsulfonamide as query fragments. As can be seen in Figure 8, **18** could only be found by a Tversky search with these fragments (X = CO-N-*, R = Et) if the chosen threshold was less than 0.65. However, choosing such a low threshold will most likely result in a large number of hits. To explore the influence of the query fragments on the outcome we have modified the choice of the core as well as the decoration of one of the query fragments. If the amide is considered to be part of the core (X = *, R = Et), the Tversky threshold increases to (a more acceptable) 0.80. If the ethyl substituent is considered to be nonessential and is removed from the query fragment (R = *, R = H), the Tversky threshold increases to 1.00. This nicely illustrates how the choice of core as well as the decoration of the query fragments will influence the likelihood of success as well as the number of hits[45] that a search will retrieve. Since these are opposing forces it will be up to the user to try and strike the right balance. It goes without saying that this might be trivial in retrospect but can otherwise be a challenge. In some instances it might therefore be beneficial to perform more than one search using different query fragments and/or Tversky thresholds.

For the CORUS search phenylsulfonamide (R = H, fragment 1) and 2-methoxyphenyl (X = *, fragment 2) were chosen in combination with a Tversky threshold of 0.85. This resulted in approximately 18 850 hits. In a manner similar to the approach we used for the previous examples, we applied filters that were derived using the reference molecule. The following filters reduced the number of hits to approximately 775: (a) Tversky threshold for both query fragments ≥0.8, (b) no more than 20 and 15 heavy atoms for in fragments 1 and 2, respectively, (c) at least one aromatic ring in the core, (d) no more than 2 rotatable bonds in the core, and (e) Tanimoto similarity ≤0.9 for the core relative to the core of the reference molecule (to eliminate cores that are very similar to the known core). Among the hits that remained after filtering were 18 molecules which had Kv1.5 data associated with them in the AurSCOPE Ion Channels database. Two examples are shown below. **22**[46] is an example for a relative conservative change in the core. This is also reflected in the fact that it has a Tanimoto coefficient of 0.69 (when compared with the reference molecule **20**) and might therefore be retrieved as a hit by a Tanimoto-based similarity search. However, it is nice to see that this molecule is among the filtered hits considering that it contains 4-trifluoromethoxyphenyl instead of 3-methoxyphenyl as one of the terminal fragments. In **21**,[47] the terminal fragment that matches the 3-methoxyphenyl query fragment is also a para-substituted phenyl ring. Encouragingly, the core of **21** is already quite different from the core of the reference molecule **20** (even though they contain similar structural elements) so that it seems justifiable to label this as a successful core hopping event. The Tanimoto coefficient for **21** (when compared with **20**) is only 0.44. For completeness, it should be mentioned that **22** as well as **21** have IC_{50} values of less than 1 μM against human Kv1.5 when tested in a patch-clamp electrophysiology assay.[46,47]

CONCLUSION

Strictly speaking, lead hopping can only be classified as core hopping if the structural difference between the query molecule and bioactive hit is limited to the central core. However, such an approach has the disadvantage that when, for example, a physically existing compound file is screened in silico only a small number of hits (if any) might be retrieved. On the other hand, if also molecules with similar terminal fragments are being considered the number of hits will increase together with the chance to identify an



alternative core. Particularly in cases where only limited SAR is available, and it is therefore not established that the terminal fragments of a given reference molecule are optimal, it seems more than reasonable to not only consider exact matches. To perform core hopping, methods are best suited that are able to recognize localized similarities despite the fact that two molecules might have significantly (localized) differences in other parts of their structures. A similarity index that is very useful in this context is the Tversky coefficient. However, as well as molecules of interest, a search with the Tversky index also retrieves a large number of false positives (in the context of core hopping). The tool we have developed can be used to postprocess hits from a similarity search by splitting them into fragments which are subsequently annotated in a way that helps the user to filter out false positives and focus on molecules of interest by means of interactive visualization. To validate our approach, we have selected four biological targets for which lead hopping examples have been published and used known actives to perform a similarity search with our tool CORUS. For the Bradykinin example, we were able to reduce the number of initial hits from 11 500 to 470 (i.e., approximately 4%) and, at the same time, successfully identify a group of four alternative cores. For the COX-2 example, the filtering led to an even more drastic reduction of hits from initially 12 700 to only 85. Very encouragingly, we found 27 molecules with reported COX-2 activity among the 85 and were able to identify a number of alternative cores that are reported in the literature. For the CB_1 example, we were still left with 3350 hits after having applied a set of filters to the 25 300 initial hits. Although it might at first seem surprising that no further reduction of the number of hits was achieved, this is most likely the result of the amount of work that has been done over the years around Rimonabant, which we have used as reference molecule. This assumption is backed up by the fact that there were 1200 molecules with CB_1 data in the AurSCOPE GPCR database among the 3350 hits. Again, it was very encouraging that we were able to identify several alternative cores among the filtered hits from the CORUS search. The same is true for the Kv1.5 example, where we were able to find alternative cores after the 18 850 hits from the initial similarity search were reduced by interactive filtering to 775. It should be particularly noted how valuable the availability of SAR knowledge bases[48] (like the AurSCOPE databases) has been for this study. Without them it would have been very much more difficult to put this approach to the test.

In summary, it can be said that by using the approach presented here we were able to identify alternative cores in all the four cases we have looked at. Although this is extremely encouraging, we acknowledge that the kind of analysis we have performed can not provide a true reflection of the potential of this approach. Only the application of the approach implemented in CORUS to ongoing drug discovery efforts will show if it will be a worthwhile addition to the ligand-based hit generation toolbox.

ACKNOWLEDGMENT

The author thanks Ceara Rea, Hilary Southgate, Martin Saunders, and Harkamal Tumber for their help in regards to Oracle and Daylight related issues. The availability of the SMILES Visualizer developed in house was crucial for the interactive visualization of the hits. For this we have to thank Gianpaolo Bravi. Also, many thanks go to Jameed Hussain, Iain McLay, Andrew Leach, and Val Gillet for most helpful discussions.

REFERENCES AND NOTES

(1) Böhm, H.-J.; Flohr, A.; Stahl, M. Scaffold Hopping. *Drug Discovery Today Technol.* **2004**, *1*, 217–224.
(2) Schneider, G.; Schneider, P.; Renner, S. Scaffold-Hopping: How Far Can You Jump? *QSAR Comb. Sci.* **2006**, *25*, 1162–1171.
(3) 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.
(4) Green, D. V. S. Virtual Screening of Virtual Libraries. *Prog. Med. Chem.* **2003**, *41*, 61–97.
(5) Shelley, M.; Frye, L. L.; Sherman, B. W.; Rao, S. N.; Beard, H.; Mozziconacci, J.-C.; Shenkin, P. S. *New Approach to Lead Optimization and Core Hopping.*, 234th ACS National Meeting, Boston, MA, August 19−23, 2007; American Chemical Society: Washington, DC, 2007; COMP-406.
(6) Feng, D.-M.; Wai, J. M.; Kuduk, S. D.; Ng, C.; Murphy, K. L.; Ransom, R. W.; Reiss, D.; Chang, R. S. L.; Harrell, C. M.; MacNeil, T.; Tang, C.; Prueksaritanont, T.; Freidinger, R. M.; Pettibone, D. J.; Bock, M. G. 2,3-Diaminopyridine As a Platform for Designing Structurally Unique Nonpeptide Bradykinin B_1 Receptor Antagonists. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 2385–2388.
(7) Kuduk, S. D.; Di Marco, C. N.; Chang, R. K.; Wood, M. R.; Schirripa, K. M.; Kim, J. J.; Wai, J. M. C.; DiPardo, R. M.; Murphy, K. L.; Ransom, R. W.; Harrell, C. M.; Reiss, D. R.; Holahan, M. A.; Cook, J.; Hess, J. F.; Sain, N.; Urban, M. O.; Tang, C.; Prueksaritanont, T.; Pettibone, D. J.; Bock, M. G. Development of Orally Bioavailable and CNS Penetrant Biphenylaminocyclopropane Carboxamide Bradykinin B_1 Receptor Antagonists. *J. Med. Chem.* **2007**, *50*, 272–282.
(8) Kuduk, S. D.; Wood, M. R.; Bock, M. G. PCT Int. Appl. WO 2004/019868, 2004.
(9) Daylight Chemical Information Systems. http://www.daylight.com (accessed April 23, 2009).
(10) Where not otherwise noted, all Daylight fingerprints used in this study have a length of 1024 bits and have been calculated using a minimal path length of zero and a maximum path length of 7.
(11) Tversky, A. Features of Similarity. *Psychol. Rev.* **1977**, *84*, 327–352.
(12) Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 2003.
(13) The Tversky similarity (S_{Tversky})[11] for two molecules A and B is calculated as follows: $S_{\text{Tversky}} = c/(\alpha(a-c) + \beta(b-c) + c)$. For binary data, a is defined as the number of bits set to "1" in molecule A. Similarly, b is defined as the number of bits set to "1" in molecule B. c is the number of bits set to "1" in both A and B. α and β are user-defined constants.
(14) Rupp, M.; Schneider, P.; Schneider, G. Distance Phenomena in High-Dimensional Chemical Descriptor Spaces: Consequences for Similarity-Based Approaches *J. Comput. Chem.* [Online early access]. DOI: 10.1002/jcc.21218 Published Online: March 5, 2009.

(15) Wood, M. R.; Schirripa, K. M.; Kim, J. J.; Wan, B.-L.; Murphy, K. L.; Ransom, R. W.; Chang, R. S. L.; Tang, C.; Prueksaritanont, T.; Detwiler, T. J.; Hettrick, L. A.; Landis, E. R.; Leonard, Y. M.; Krueger, J. A.; Lewis, S. D.; Pettibone, D. J.; Freidinger, R. M.; Bock, M. G. Cyclopropylamino Acid Amide as a Pharmacophoric Replacement for 2,3-Diaminopyridine. Application to the Design of Novel Bradykinin B_1 Receptor Antagonists. *J. Med. Chem.* **2006**, *49*, 1231–1234.
(16) Babu, Y. S.; Rowland, R. S.; Chand, P.; Kotian, P. L.; El-Kattan, Y.; Niwas, S. U.S. Patent 6,699,994, 2004.
(17) Wagener, M.; Lommerse, J. P. M. The Quest for Bioisosteric Replacements. *J. Chem. Inf. Model.* **2006**, *46*, 677–685.
(18) Leach, A. R.; Green, D. V. S.; Hann, M. M.; Harper, G.; Whittington, A. R. *SIV: A Synergistic Approach to the Analysis of High-Throughput Screening Data*; 221st ACS National Meeting, San Diego, CA, April 1−5, 2001; American Chemical Society: Washington, DC, 2001; ClNF-080.
(19) CORUS is an acronym and stands for core replacement utility script.
(20) Aureus Pharma. http://www.aureus-pharma.com (accessed April 23, 2009).
(21) Arsenic has been chosen as a "dummy atom" in the fragmentation algorithm since it normally does not occur in the molecules we are interested in and at the same time is one of the view elements that can be a member of an aromatic ring in a SMILES string. The charge of the arsenic atom is set according to the valency of the atomic position.
(22) There are two parameters in the fragmentation algorithm that determine how many mutants are considered for further analysis. The first parameter is multiplied with the Tversky similarity value for the top match and the result is used as threshold. For this study, the parameter was set to 0.95. If for example, the maximum Tversky similarity value is 1.0, this would mean that all mutants with a Tversky similarity value of 0.95 or more will be considered in the following steps. However, to avoid a case where a large number of mutants will have to be looked at, we have also introduced a second parameter that restricts the number of mutants that will be considered (and that in this study was set to 10).
(23) Bayada, D. M.; Simpson, R. W.; Johnson, A. P.; Laurenco, C. An algorithm for the multiple common subgraph problem. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 680–685.
(24) TIBCO. http://spotfire.tibco.com (accessed April 23, 2009).
(25) Wood, M. R.; Su, D.-S.; Wai, J. M.-C. U.S. Patent 2006/0173023, 2006.
(26) Gauthier, J. Y.; Leblanc, Y.; Black, W. C.; Chan, C.-C.; Cromlish, W. A.; Gordon, R.; Kennedy, B. P.; Lau, C. K.; Léger, S.; Wang, Z.; Ethier, D.; Guay, J.; Mancini, J.; Riendeau, D.; Tagari, P.; Vickers, P.; Wong, E.; Xu, L.; Prasit, P. Synthesis and Biological Evaluation of 2,3-Diarylthiophenes As Selective COX-2 Inhibitors. Part II: Replacing the Heterocycle. *Bioorg. Med. Chem. Lett.* **1996**, *6*, 87–92.
(27) Prasit, P.; Wang, Z.; Brideau, C.; Chan, C.-C.; Charleson, S.; Cromlish, W.; Ethier, D.; Evans, J. F.; Ford-Hutchinson, A. W.; Gauthier, J. Y.; Gordon, R.; Guay, J.; Gresser, M.; Kargman, S.; Kennedy, B.; Leblanc, Y.; Léger, S.; Mancini, J.; O'Neill, G. P.; Ouellet, M.; Percival, M. D.; Perrier, H.; Riendeau, D.; Rodger, I.; Tagari, P.; Thérien, M.; Vickers, P.; Wong, E.; Xu, L.-J.; Young, R. N.; Zamboni, R.; Boyce, S.; Rupniak, N.; Forrest, M.; Visco, D.; Patrick, D. The Discovery of Rofecoxib [MK 966, VIOXX,4-(4′-Methylsulfonylphenyl)-3-phenyl-2(5H)-furanone], An Orally Active Cyclooxygenase-2 Inhibitor. *Bioorg. Med. Chem. Lett.* **1999**, *9*, 1773–1778.
(28) Friesen, R. W.; Brideau, C.; Chan, C. C.; Charleson, S.; Deschênes, D.; Dubé, D.; Ethier, D.; Fortin, R.; Gauthier, J. Y.; Girard, Y.; Gordon, R.; Greig, G. M.; Riendeau, D.; Savoie, C.; Wang, Z.; Wong, E.; Visco, D.; Xu, L. J.; Young, R. N. 2-Pyridinyl-3-(4-methylsulfonyl)phenylpyridines: Selective and Orally Active Cyclooxygenase-2 Inhibitors. *Bioorg. Med. Chem. Lett.* **1998**, *8*, 2777–2782.
(29) Penning, T. D.; Talley, J. J.; Bertenshaw, S. R.; Carter, J. S.; Collins, P. W.; Docter, S.; Graneto, M. J.; Lee, L. F.; Malecha, J. W.; Miyashiro, J. M.; Rogers, R. S.; Rogier, D. J.; Yu, S. S.; Anderson, G. D.; Burton, E. G.; Cogburn, J. N.; Gregory, S. A.; Koboldt, C. M.; Perkins, W. E.; Seibert, K.; Veenhuizen, A. W.; Zhang, Y. Y.; Isakson, P. C. Synthesis and Biological Evaluation of the 1,5-Diarylpyrazole Class of Cyclooxygenase-2 Inhibitors: Identification of 4-[5-(4-Methylphenyl)-3-(trifluoromethyl)-1H-pyrazol-1-yl]benzenesulfonamide (SC-58635, Celecoxib). *J. Med.Chem.* **1997**, *40*, 1347–1365.
(30) Talley, J. J.; Brown, D. L.; Carter, J. S.; Graneto, M. J.; Koboldt, C. M.; Masferrer, J. L.; Perkins, W. E.; Rodgers, R. S.; Shaffer, A. F.; Zhang, Y. Y.; Zweifel, B. S.; Seibert, K. 4-[5-Methyl-3-phenylisoxazol-4-yl]-benzenesulfonamide, Valdecoxib: A Potent and Selective Inhibitor of COX-2. *J. Med. Chem.* **2000**, *43*, 775–777.
(31) Lohray, B. B.; Lohray, V. B.; Jain, M. R.; Patel, G. D.; Pingali, H. PCT Int. Appl. WO 2003/087062, 2003.
(32) Puig Duran, C.; Feixas Gras, J.; Jimenez Mayorga, J. M.; Crespo Crespo, M. I. PCT Int. Appl. WO 1999/14205 1999.
(33) Reddy, M. V. R.; Bell, S. C. PCT Int. Appl. WO 2003/024958, 2003.
(34) Rinaldi-Carmona, M.; Barth, F.; Héaulme, M.; Shire, D.; Calandra, B.; Congy, C.; Martinez, S.; Maruani, J.; Néliat, G.; Caput, D.; Ferrara, P.; Soubrié, P.; Brelière, J. C.; Le Fur, G. SR141716A, A Potent and Selective Antagonist of the Brain Cannabinoid Receptor. *FEBS Lett.* **1994**, *350*, 240–244.
(35) Lange, J. H. M.; Kruse, C. G. Medicinal Chemistry Strategies to CB_1 Cannabinoid Receptor Antagonists. *Drug Discovery Today* **2005**, *10*, 693–702.
(36) Berggren, A. I. K.; Bostrom, S. J.; Cheng, L.; Elebring, S. T.; Greasley, P.; Nagard, M.; Wilstermann, J. M.; Terricabras, E. PCT Int. Appl. WO 2004/058249, 2004.
(37) Ellsworth, B. A.; Wang, Y.; Zhu, Y.; Pendri, A.; Gerritz, S. W.; Sun, C.; Carson, K. E.; Kang, L.; Baska, R. A.; Yang, Y.; Huang, Q.; Burford, N. T.; Cullen, M. J.; Johnghar, S.; Behnia, K.; Pelleymounter, M. A.; Washburn, W. N.; Ewing, W. R. Discovery of pyrazine carboxamide CB1 antagonists: The introduction of a hydroxyl group improves the pharmaceutical properties and in vivo efficacy of the series. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 3978–3982.
(38) Toupence, R. B.; Debenham, J. S.; Goulet, M. T.; Madsen-Duggan, C. B.; Walsh, T. F.; Shah, S. K. PCT Int. Appl. WO 2004/012671, 2004.
(39) Moritani, Y.; Shirai, K.; Oi, M. PCT Int. Appl. WO 2007/046548, 2007.
(40) Alanine, A.; Bleicher, K.; Guba, W.; Haap, W.; Kube, D.; Luebbers, T.; Plancher, J.-M.; Roche, O.; Rogers-Evans, M.; Schneider, G.; Zuegge, J. U.S. Patent 2004/0142922, 2004.
(41) Lange, J. H. M.; Kruse, C. G.; van Stuivenberg, H. H. U.S. Patent 2005/0171179, 2005.
(42) Carpino, P. A.; Dow, R. L. U.S. Patent Appl. US 2004/02114855, 2004.
(43) Schneider, G.; Clément-Chomienne, O.; Hilfiger, L.; Schneider, P.; Kirsch, S.; Böhm, H.-J.; Neidhart, W. Virtual Screening for Bioactive Molecules by Evolutionary De Novo Design. *Angew. Chem., Int. Ed.* **2000**, *39*, 4130–4133.
(44) Castle, N. A.; Hollinshead, S. P.; Hughes, P. F.; Mendoza, J. S.; Wilson, J. W.; Amato, G.; Beaudoin, S.; Gross, M.; McNaughton-Smith, G. PCT Int. Appl. WO 1998/04521, 1998.
(45) To get a better feeling for how the choice of the Tversky threshold and query fragments impacts on the number of hits that are retrieved we run three searches with the combinations listed in Figure 8. The number of hits retrieved from the AurScope databases available to us are the following: (a) 104 517 hits for a Tversky threshold of 0.646 and X = CO-N-*/R = Et, (b) 45 438 hits for a Tversky threshold of 0.719 and X = CO-N-*/R = H, and (c) 23 789 hits for a Tversky threshold of 0.789 and X = */R = Et.
(46) Gross, M.; Castle, N. A. Int. PCT Appl. WO 1999/37607, 1999.
(47) Brendel, J.; Pirard, B.; Peukert, S.; Kleeman, H.-W.; Hemmerle, H. U.S. Patent Appl. US 2007/0117807, 2007.
(48) Senger, S.; Leach, A. R. SAR Knowledge Bases in Drug Discovery. *Ann. Rep. Comput. Chem.* **2008**, *4*, 203–216.

CI900092Y