

Predicting Key Example Compounds in Competitors' Patent Applications Using Structural Information Alone

Kazunari Hattori,^{*,†,§} Hiroaki Wakabayashi,[‡] and Kenta Tamaki[‡]

Medicinal Chemistry Technologies and Research Informatics, Pfizer Global Research and Development, Nagoya Laboratories, Pfizer Inc., 5-2 Taketoyo, Aichi 470-2393, Japan

Received July 25, 2007

In drug discovery programs, predicting key example compounds in competitors' patent applications is important work for scientists working in the same or in related research areas. In general, medicinal chemists are responsible for this work, and they attempt to guess the identity of key compounds based on information provided in patent applications, such as biological data, scale of reaction, and/or optimization of the salt form for a particular compound. However, this is sometimes made difficult by the lack of such information. This paper describes a method for predicting key compounds in competitors' patent applications by using only structural information of example compounds. Based on the assumption that medicinal chemists usually carry out extensive structure–activity relationship (SAR) studies around key compounds, the method identifies compounds located at the centers of densely populated regions in the patent examples' chemical space, as represented by Extended Connectivity Fingerprints (ECFPs). For the validation of the method, a total of 30 patents containing structures of launched drugs were selected to test whether or not the method is able to predict key compounds (the launched drugs). In 17 out of the 30 patents (57%), the method was able to successfully predict the key compounds. The result indicates that our method could provide an alternative approach to predicting key compounds in cases where the conventional medicinal chemist's approach does not work well. This method could also be used as a complement to the traditional medicinal chemist's approach.

INTRODUCTION

Patents are legal rights for applicants to protect their inventions for a fixed period of time.¹ Therefore, a large amount of patent applications in drug discovery programs have been filed from pharmaceutical and biotechnology companies to obtain exclusive rights on their inventions. Since patent applications describe details of inventions in claims and examples, they are one of the major sources of information for drug discovery, together with scientific literature and presentations at research conferences.

Recently, as the demand for analysis of a large amount of text data, such as patent documents and scientific literature, has been increasing, techniques for text data mining have received enormous attention.^{2–6} In the area of bioinformatics, these techniques have been used extensively to extract hidden relationships between genes, proteins, and diseases and to make novel hypotheses about biological systems.^{7–20} Patent analyses that deal with topics relevant to research planning, such as competitors' activities in a specific area of research and R&D trends in specific companies, are also areas in which text mining techniques have been employed.^{21–23} Numerous commercial tools, for example, STNAnaVist, Pipeline Pilot Text Analytic Collection, OmniViz, and ClearForest, have been made available to help researchers carry out text mining and data visualization.^{24–27} In the field

of chemoinformatics, generation of chemical structures from chemical names and/or structural images in patents and literature is the initial step necessary to perform various analyses using the structural information. Several applications to perform this step have been reported.²⁸ Since structure-searchable databases, such as MDL Patent Chemistry Database and Jubilant's databases, are now commercially available, it is possible to use these databases instead of creating new ones *de novo*.^{29,30} Once the structure database has been prepared, it can be used in a wide variety of scenarios such as prioritization of hit compounds from high-throughput screening (HTS) and comparison with the in-house compounds database. A recent article has reported that this kind of database has provided significant value to the drug discovery programs at AstraZeneca.²⁸

When analyzing a competitor's chemical patent application, retrieving as much information as possible is important for researchers working in the same or related research areas. Thus, when a new patent application is published, medicinal chemists generally read the patent document to understand claims and to identify key example compounds (in this paper, key example compounds mean the most important compounds for the applicant, not necessarily the compounds with the most potent biological activities). Predicting key compounds in the patent application is important work for the following reasons. First, in drug discovery programs, structural modification of competitors' key compounds is one of the strategies used to obtain novel lead compounds, along with other avenues, such as HTS, fragment screening, virtual screening, and structure-based design.^{31–33} Another reason is that knowing the profiles of competitors' key compounds,

* Corresponding author e-mail: kazunari.hattori@shionogi.co.jp.

† Medicinal Chemistry Technologies.

‡ Research Informatics.

§ Current address: Discovery Research Laboratories, Shionogi & Co., Ltd., 12-4, Sagisu, 5-chome, Fukushima-ku, Osaka 553-0002, Japan.

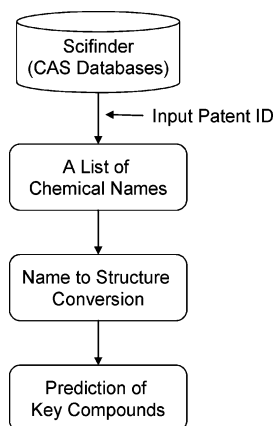


Figure 1. Overview of the process for key compounds prediction.

such as biological activities and ADMET properties, is helpful in developing a strategy for differentiation. As mentioned above, identifying key compounds in a patent application is commonly carried out by medicinal chemists based on the published information, such as biological data, scale of reaction, and/or salt screening for a particular compound. However there are often cases where this identification is made difficult by the absence or the paucity of this information.

The aim of the present paper is to introduce a computational method for predicting key compounds in a patent application by solely using the structural information on the whole set of examples. Our method is applicable to any substance patent application since it does not use any information other than the chemical structures of example compounds. The results of a validation study of the method using 30 patents which contain launched drugs are also presented.

METHOD

Our method for predicting key compounds follows the 3 major steps shown in Figure 1: the first step consists of preparing a list of chemical names of the patent example compounds; the second step consists of converting the chemical names into chemical structures; and the final step consists of predicting the key compounds using the chemical structures. Details of procedures for all three steps are described below.

Preparation of a Structure File for Patent Example Compounds. By using Scifinder³⁴ Chemical Abstracts Service (CAS), databases were searched for a patent application by entering its patent ID, and then, among substances indexed in the patent application, only those with “Biological Study” or “Preparation” roles were retrieved. Information pertaining to these substances was saved as “Quoted Format”, and CA Index Names were extracted to create a list file of these chemical names. The file prepared was used as an input for the CambridgeSoft Batch Struct=Name³⁵ program to convert the chemical names into structures. For CA Index Names which Struct=Name software failed to convert into corresponding structures, the structures were manually sketched using structure drawing software. When chemical structures contained salts, such as a hydrochloride or sodium salt, the salts were removed from the structures.

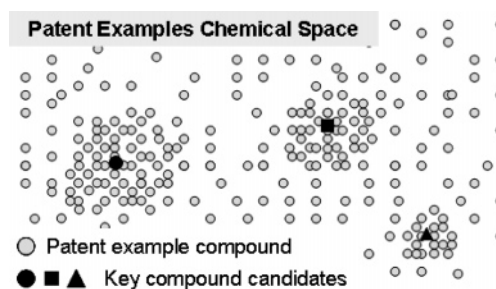


Figure 2. Graphical image of patent example compounds in chemical space. Each gray circle represents an example compound. The black circle, square, and triangle represent key compound candidates.

Computational Algorithm To Predict Key Compounds.

Before starting to develop an algorithm for key compounds prediction, we formulated the following hypothesis: “If patent example compounds were to be plotted in chemical space, they would distribute *unevenly* in the chemical space; that is, the chemical map would contain sparse areas as well as some dense areas (clusters). The clusters should contain key compounds at their center because the SAR around those important compounds must have been extensively explored by the medicinal chemists.” Figure 2 illustrates an image of example compounds spread in a representation of chemical space.

Based on the hypothesis mentioned above, a method for predicting key compounds was developed, using Scitegic Pipeline Pilot 5.0.³⁶ The method follows the procedures outlined as follows: 1. Select a compound from all example compounds and calculate similarity values (Tanimoto coefficient³⁷) with the remaining compounds. Then, count the number of compounds that have similarity values higher than a user-defined similarity cutoff. In this step, two types of fingerprints (ECFPs and MDL public keys³⁸) were used for structural descriptors to calculate similarity values. 2. Repeat step 1 for all compounds. 3. Based on the results of steps 1 and 2, identify a compound (or compounds) with the largest number of similar structures (compounds with Tanimoto similarity values higher than the user-defined cutoff value). This compound becomes the first key compound candidate. If there are multiple compounds with the same number of similar structures, then they are all regarded as key compound candidates. 4. Remove the key compound(s) identified in step 3 and their clusters (similar compounds) from the population. 5. Repeat steps 1–4 to find other key compound candidates. In this method, steps 1–4 were repeated 5 times to identify at least 5 key compound candidates. When the predicted key compounds feature several possible stereoisomers, they are regarded as one compound.

Validation of Methodology. To confirm whether or not our hypothesis mentioned above is reasonable, principal components analyses (PCA) were carried out on patent example sets using ECFP_4 fingerprints with a fixed length of 1024 bits as molecular descriptors, and distribution of example compounds was visualized in a low dimensional space. Pipeline Pilot was used to perform PCA, and fingerprints’ bits were scaled by MeanCenter.

Validation of the algorithm for key compounds prediction was performed as follows. We selected launched drugs from a list of U.S. top selling 200 drugs in 2005³⁹ on the condition that a chemical patent of the drug contain more than 50

Table 1. Patents Used for Validation of the Methodology

sales_rank	product	company	patent_ID	no. of examples
194	Levitra	Bayer	WO1999024433(A1)	330
26	Celebrex	Pfizer, Inc	WO1995015316(A1)	276
52	Aricept	Eisai Co	EP296560(A2)	265
132	Aldara	3M Pharmaceuticals	EP145340(A2)	246
60	Cozaar	Merck & Co.,Inc	EP253310(A2)	229
126	Benicar	Sankyo Pharma	EP503785(A1)	222
65	Detrol	Pfizer, Inc	EP0325571(A1)	169
138	Lescol	Novartis	WO1984002131(A1)	163
162	Casodex	Astrazeneca Pharm. Lp	EP100172(A1)	162
181	Tarceva	Genentech Inc	WO1996030347(A1)	157
113	Patanol	Nestle Sa	EP235796(A2)	151
106	Cialis	Lilly ICOS	WO1995019978(A1)	118
43	Diovan	Novartis	EP443983(A1)	118
92	Avapro	Bristol-Myers Squibb Corp	WO1991014679(A1)	107
109	Flovent	Glaxosmithkline	NL8100707(A)	104
184	Bextra	Pfizer, Inc	WO1996025405(A1)	102
32	Aciphex	Eisai Co	EP268956(A2)	100
57	Lamisil	Novartis	EP24587(A1)	91
186	Vigamox	Alcon Inc	EP550903(A1)	88
154	Femara	Novartis	EP236940(A2)	86
189	Zomig	Medpointe Pharm	WO1991018897(A1)	77
64	Zofran	Glaxosmithkline	DE3502508(A1)	74
97	Spiriva	Boehringer Ingelheim	EP418716(A1)	72
41	Coreg	Glaxosmithkline	DE2815926(A1)	71
183	Atacand	Astrazeneca	EP459136(A1)	67
98	Paxil	Glaxosmithkline	EP266574(A2)	63
55	Nasonex	Schering-Plough Corp.	EP57401(A1)	61
103	Sustiva	Bristol-Myers Squibb Corp	EP582455(A1)	59
93	Arimidex	Astrazeneca Pharm. Lp	EP296749(A1)	56
101	Reyataz	Bristol-Myers Squibb Corp	WO1997040029(A1)	55

example compounds in it and tested whether or not the method was able to successfully predict the structures of the drugs as key compounds. If the structure of the drug appears within the top 5 predicted key compounds, then we considered that the method could correctly predict the key compound. Scifinder was utilized to search for chemical patents of drugs. Among patents retrieved by this search, the earliest chemical patent associated with the drug was chosen for the analysis. We regarded the number of substances indexed in the patent in CAS databases as the number of example compounds. Table 1 lists the information on the 30 drugs employed in this validation study, and Figure 3 shows the drug structures. As can be seen, there are no biases in terms of companies, number of examples, or chemical structures.

Fingerprints for Similarity Calculation. Two different types of fingerprints, ECFPs and MDL public keys, were employed in step 1 of the key compounds prediction scheme. ECFPs are circular substructure fingerprints which are generated by using a variant of Morgan algorithm.⁴⁰ The initial code is assigned to an atom in a molecule based on the number of connections to the atom, the element type, the charge, and the mass of the atom. This initial code represents structural information about the atom itself. In the next iteration, a new code is generated by using the codes from immediate neighbor atoms to represent a larger structural feature around the atom. This iteration process is repeated until it reaches the maximum diameter of the neighborhoods a user defines. For example, ECFP₄ generates structural features around each atom up to a diameter of 4 in bonds, which is achieved by two iterations. MDL public keys, on the other hand, are binary descriptors which encode the presence or absence of 166 predefined substructure queries. These were generated by using a Pipeline Pilot component.

RESULTS AND DISCUSSION

Distribution of Patent Examples on PCA Plots. Figure 4 shows the results of PCA for WO1995015316(A1) and EP296749(A1), which are chemical patents that include structures of Celebrex (Pfizer) and Arimidex (Astrazeneca), respectively. We selected these two patents because WO1995015316(A1) has a large number of example compounds ($N = 276$) and EP296749(A1) has a small number of example compounds ($N = 56$) among the 30 patents used for the validation study. The plot of WO1995015316(A1) clearly shows that there are several densely populated regions. Even EP296749(A1), with a much smaller number of example compounds, had a highly dense area and other sparse areas. Since PCs 1–3 account for only 40% and 60% of total variances for WO1995015316(A1) and EP296749(A1), respectively, the black circles do not express accurate positions of Celebrex and Arimidex in high dimensional chemical spaces. However, it would be certain that both Celebrex and Arimidex are located in dense areas. These results imply that our hypothesis is reasonable enough to be used as the basis for further validation study.

Investigation of an Optimal Combination of Fingerprints and a Similarity Cutoff Value. To determine the fingerprints and similarity cutoff value which give the best performance in terms of prediction accuracy, all combinations of ECFPs (ECFP₂, ECFP₄, and ECFP₆) and similarity cutoff values (0.6, 0.7, and 0.8), a total of 9 parameter sets, were applied to the key compounds prediction. As for MDL public keys, similarity cutoff values of 0.9 and 0.95 (a total of 2 parameter sets) were used since our preliminary analysis revealed that MDL public keys generally provide higher similarity values than ECFPs for a set of reference and target compounds (data not shown). Table 2 summarizes the results. The use of MDL public keys showed poor prediction

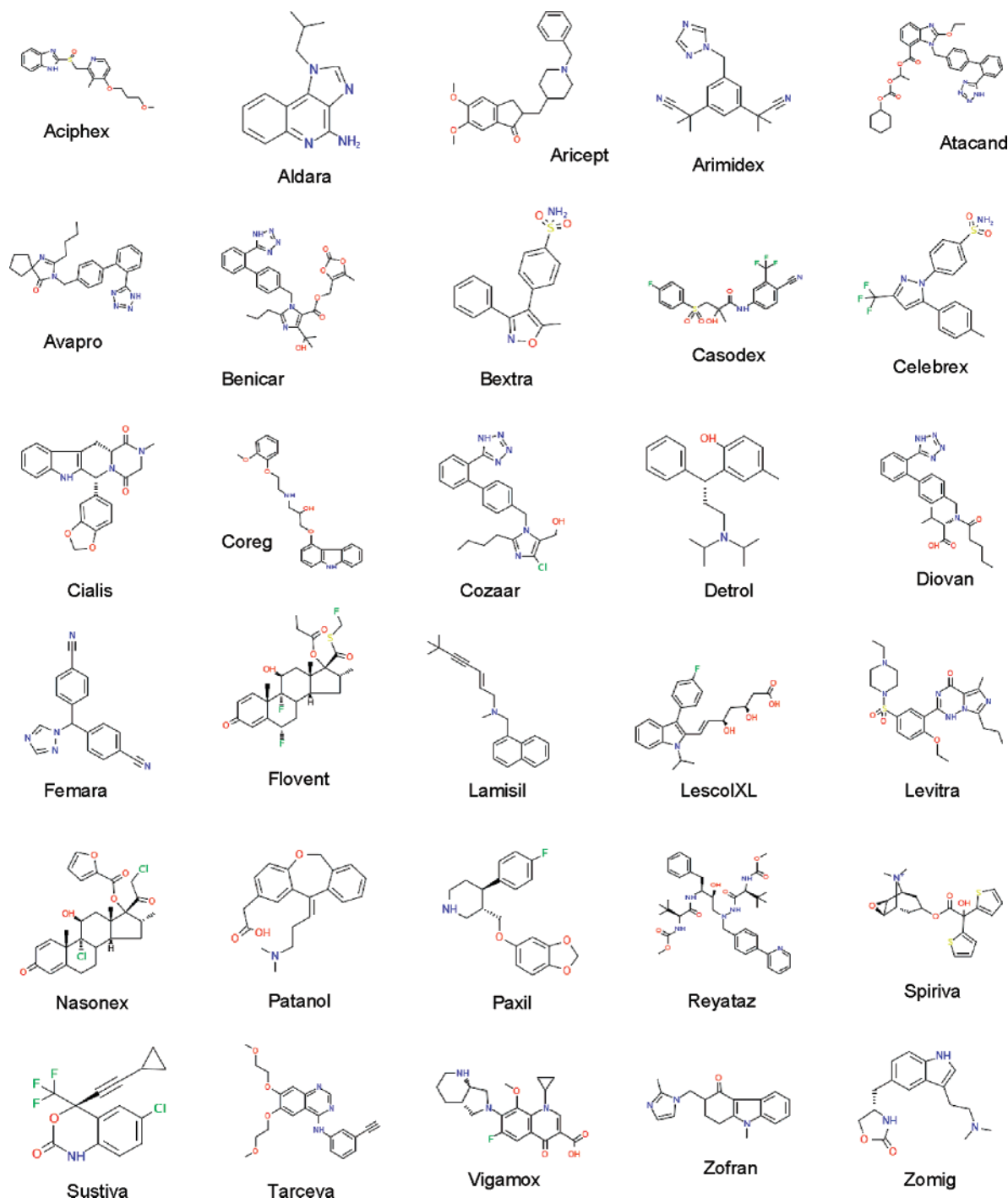


Figure 3. Structures of drugs used for validation study.

accuracy regardless of similarity cutoff values. Regarding ECFPs, on average, ECFP_6 fingerprints showed a prediction accuracy higher than those of ECFP_2 and ECFP_4. The order of prediction accuracy was ECFP_6 > ECFP_4 > ECFP_2, and the average accuracy rate was 42% for ECFP_6. As for the similarity cutoff value, a value of 0.8 outperformed 0.6 and 0.7. The use of a similarity cutoff of 0.8 gave a prediction accuracy of 43% on average. However, when we look at individual cases, the highest prediction accuracy was obtained when the combination of ECFP_4 and the similarity cutoff value of 0.7 was used. In this case, the method was able to predict key compounds correctly in 17 out of 30 patents. This corresponds to 57% prediction accuracy. The combination of ECFP_6 and the similarity

cutoff value of 0.8, which was expected to show the best performance, gave the second-best result (50%).

Since ECFPs encode structural information around each atom in a molecule, they are sensitive to minor differences in molecular structures. MDL public keys, on the other hand, encode only the presence or absence of a small set of predefined structural queries. Thus, they would be less sensitive to such minor differences of structures. In general, example compounds in a patent are structurally similar; therefore, the capability to detect small differences in structures would be an important factor for fingerprints used in this patent analysis.

Table 3 shows individual results for the validation studies. Among actual key compounds, only Arimidex was success-

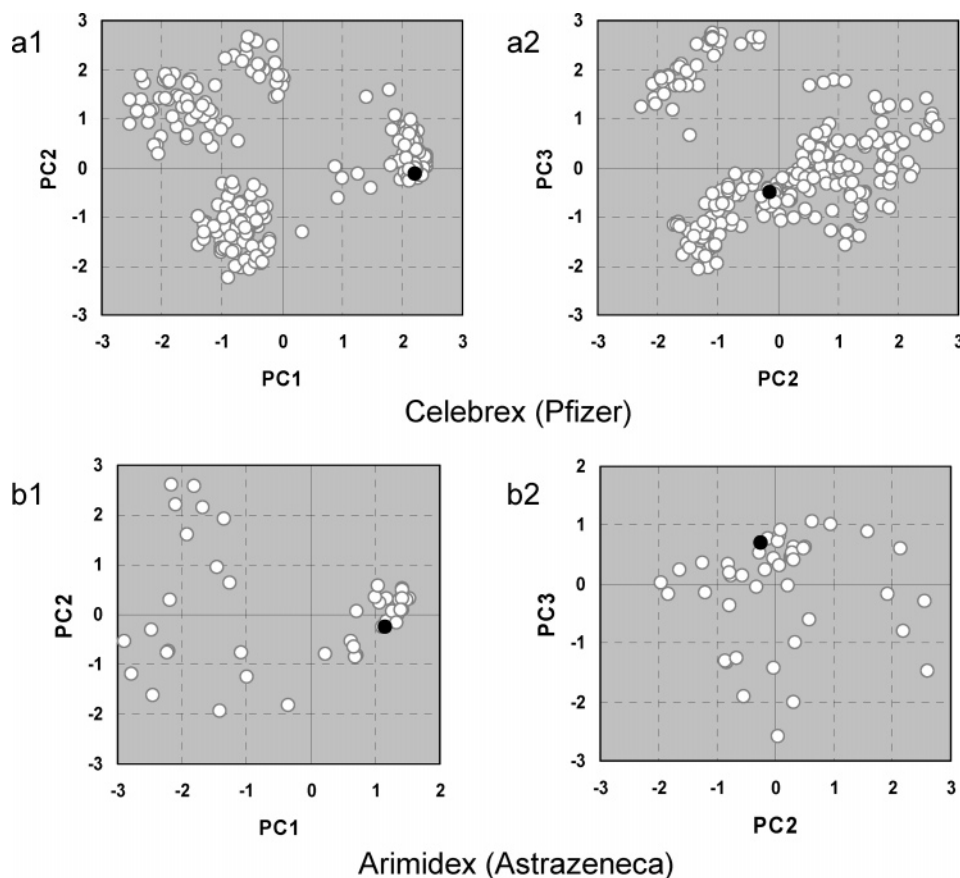


Figure 4. PCA plots for WO1995015316(A1) and EP296749(A1). The plots, a1 and b1, are PC1 vs PC2, and the plots, a2 and b2, are PC2 vs PC3 for WO1995015316(A1) and EP296749(A1), respectively. The open circles mean example compounds, and the filled circles mean drugs (Celebrex and Arimidex).

Table 2. Summary Table for the Results of the Validation Study

fingerprints	similarity cutoff					total
	0.6	0.7	0.8	0.9	0.95	
ECFP_2	7 (0.23) ^a	8 (0.27)	13 (0.43)			28 (0.31)
ECFP_4	6 (0.20)	17 (0.57)	11 (0.37)			34 (0.38)
ECFP_6	11 (0.37)	12 (0.40)	15 (0.50)			38 (0.42)
MDL keys				6 (0.20)	8 (0.27)	
total	24 (0.27)	37 (0.41)	39 (0.43)			

^a Number of patents whose key compounds were predicted successfully by the method. The figure in parentheses is the accuracy rate of prediction.

fully predicted in all of the parameter sets, and for 5 drugs (Celebrex, Cozaar, Vigamox, Nasonex, Reyataz), only one parameter set gave the correct answer. This result indicates that the selection of fingerprints and similarity cutoff value greatly affects the performance of the prediction. On the other hand, for 25 out of 30 drugs (83%), actual key compounds were successfully predicted in at least one of the 11 parameter sets. This result indicates that our initial hypothesis (that actual key compounds are located at the center of highly populated areas) is borne out as long as the selection of fingerprints and cutoff values is appropriate. Specifically, the combination of ECFP_4 and the similarity cutoff value of 0.7 can be used practically for prediction due to its relatively good reliability (57%).

Table 4 summarizes the drugs' positions among the top 5 predicted key compounds, when ECFP_4 fingerprints and a similarity cutoff value of 0.7 were used. It is worth noting that about 60% of drugs appeared in the first rank, but the

remaining 40% of drugs were ranked second or third. This means that actual key compounds are not always located at the center of the most densely populated region of the example compounds' chemical space. This may be due to the fact that an applicant often adds examples for 1 year after filing a first patent application by using internal priority rights.¹ In general, in the early stage of the drug discovery process, medicinal chemists extensively explore the SAR of their lead structures in a systematic manner; afterward, they continue optimization more efficiently by making use of the SAR information they have gathered. Therefore, it would not be uncommon to find more attractive compounds after filing a patent application and add them to the original examples set to make the patent position stronger. Considering all these factors, it is important to see not only the top 1 predicted key compound but also lower order compounds. The result in Table 4 indicates that considering all top 3 predicted key compounds is necessary so as not to miss any actual key compounds.

Investigation of Relationship between Molecular Size and Prediction Accuracy. As is reported in the literature, Tanimoto coefficients calculated by using 2D fragment descriptors is sensitive to the size of molecules.^{41,42} In general, similarity calculations within small size molecules produce lower average Tanimoto coefficients than those within large size molecules. As is shown in Table 3, our validation patent set is diverse in terms of average molecular weight, ranging from 260 to 710. Thus, we investigated whether or not there are any trends in the accuracy rate for

Table 3. Results of Validation Study for Each Drug

drugs	MW_av ^a	ECFP_2			ECFP_4			ECFP_6			MDL keys		consensus ^c
		0.6 ^b	0.7	0.8	0.6	0.7	0.8	0.6	0.7	0.8	0.9	0.95	
Levitra	531.4	×	×	×	×	×	×	×	×	o	×	o	o
Celebrex	407.3	×	×	×	×	o	×	×	×	×	×	×	o
Aricept	373.3	×	o	o	o	o	o	o	o	o	×	o	o
Aldara	265.2	×	×	o	o	×	o	×	o	o	o	o	o
Cozaar	470.7	o	×	×	×	×	×	×	×	×	×	×	o
Benicar	476.5	×	×	×	×	×	×	×	×	×	×	×	×
Detrol	357.6	o	×	×	×	×	×	×	×	o	×	×	o
Lescol	415.6	×	×	o	×	o	×	o	×	o	o	o	o
Casodex	359.4	×	×	×	×	×	×	×	×	×	×	×	×
Tarceva	368.0	×	×	o	×	o	o	o	o	o	o	×	o
Patanol	353.6	×	o	×	×	o	o	×	o	×	×	o	o
Cialis	417.5	×	o	×	×	o	×	o	×	o	×	×	o
Diovan	461.7	×	×	o	×	o	×	×	o	o	o	×	o
Avapro	385.4	×	×	×	×	×	×	×	o	o	×	×	o
Flovent	488.7	×	×	o	×	o	o	o	×	o	×	×	o
Bextra	359.9	×	o	×	o	o	×	×	×	×	o	×	o
Aciphex	401.7	×	×	o	×	o	×	o	×	o	×	×	o
Lamisil	273.6	×	×	o	×	o	o	o	×	×	×	×	o
Vigamox	439.0	o	×	×	×	×	×	×	×	×	×	×	o
Femara	259.1	×	×	×	×	×	×	×	×	×	×	×	×
Zomig	271.2	×	o	o	×	o	o	o	o	o	×	×	o
Zofran	334.9	o	o	o	o	o	×	o	o	o	×	×	o
Spiriva	384.7	×	×	×	×	×	o	×	o	×	×	o	o
Coreg	371.8	o	o	o	o	o	o	o	o	×	×	o	o
Atacand	480.8	×	×	×	×	×	×	×	×	×	×	×	×
Paxil	392.7	×	×	×	×	×	×	×	×	×	×	×	×
Nasonex	532.6	o	×	×	×	×	×	×	×	×	×	×	o
Sustiva	309.5	×	×	o	×	o	o	×	o	o	×	×	o
Arimidex	297.0	o	o	o	o	o	o	o	o	o	o	o	o
Reyataz	710.1	×	×	×	×	o	×	×	×	×	×	×	o
total		7	8	13	6	17	11	11	12	15	6	8	25

^a Average molecular weight of example compounds. ^b Similarity cutoff value. ^c If a key compound was successfully predicted in any of the 11 parameter sets, the judgment becomes o, if not ×.

Table 4. Rankings of Marketed Drugs within Top 5 Predicted Key Compounds

rank	no. of drugs	drug names
1	10	Aricept, Arimidex, Celebrex, Cialis, Coreg, Diovan, Patanol, Tarceva, Zofran, Zomig
2	4	Aciphex, Flovent, Lamisil, Lescol
3	3	Bextra, Reyataz, Sustiva
4	0	
5	0	

each average molecular size range. This could enable us to set a similarity cutoff value with a high probability of successful prediction when information on the average molecular weight of the example compounds is provided.

Figure 5 summarizes the prediction results for each binned average molecular weight. For patents with compounds of average molecular weight between 200 and 300, the number of patents whose key compounds were successfully predicted was increased as the similarity cutoff value became higher. No significant differences were observed for patents where the average molecular weight was between 300 and 400. What is most notable is that the use of a similarity cutoff value of 0.7 showed much better prediction accuracy than the other two values did for patents with average molecular weight between 400 and 500. For patents with average molecular weight over 500, prediction accuracy was very poor for all similarity cutoff values.

These results suggest that the use of a similarity cutoff value of 0.7 combined with ECFP_4 fingerprints is a good

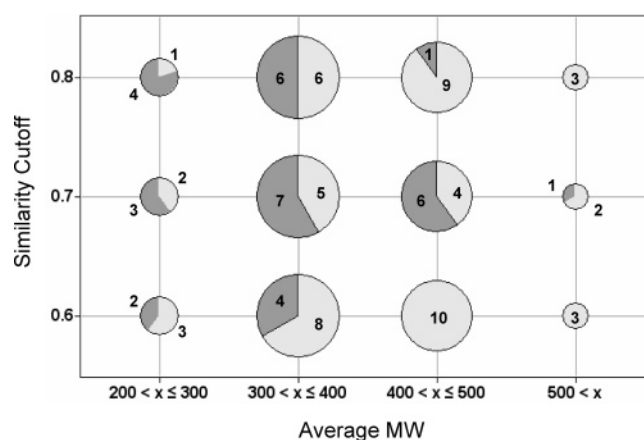


Figure 5. Pie charts showing prediction results for subsets grouped by average molecular weight of example compounds and similarity cutoff values. The results using ECFP_4 fingerprints are shown in the charts. Gray and pale gray portions represent the proportion of success and failure of prediction respectively. Numbers in pie charts mean the number of patents. Size of each pie chart represents the number of patents belonging to each subset.

choice especially for a patent application with compounds whose average molecular weight is between 300 and 500. As for small size compounds with an average molecular weight of ≤ 200 and large size compounds with an average molecular weight of > 500 , we need further validation studies with larger patent sets to come to a conclusion.

Prediction Output of the Analysis. Figure 6 shows a typical prediction output, generated by the Pipeline Pilot protocol. The result for WO1995015316(A1) (Celebrex) is

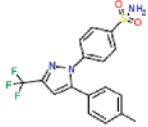

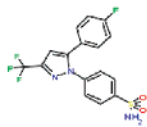

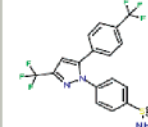

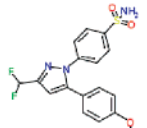
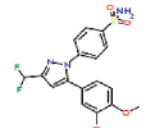
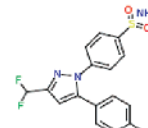
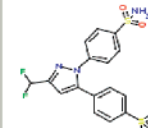
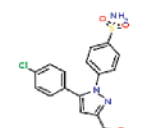
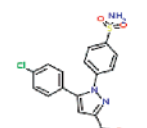
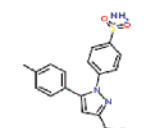
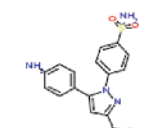
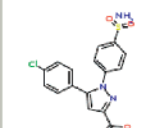
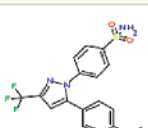
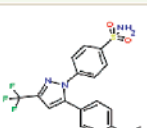
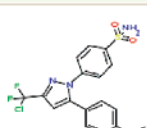
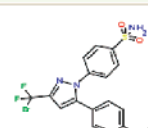
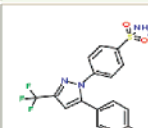
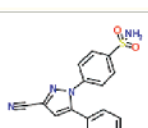
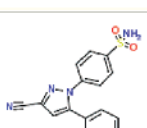
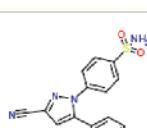
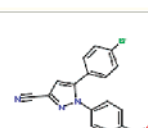
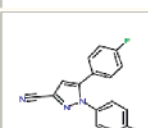
Molecule	Name	Canonical Smiles	NumberOfClosest	Similar_Compounds1	Similar_Compounds2	Similar_Compounds3	Similar_Compounds4
	Benzenesulfonamide, 4-[3-(4-methylphenyl)-3-(trifluoromethyl)-1H-pyrazol-1-yl]-	<chem>Cc1ccc(cc1)c2cc(m2c3ccc(cc3)S(=O)(=O)N)C(F)(F)F</chem>	33				
	Benzenesulfonamide, 4-[3-(difluoromethyl)-5-(4-methoxyphenyl)-1H-pyrazol-1-yl]-	<chem>COc1ccc(cc1)c2cc(m2c3ccc(cc3)S(=O)(=O)N)C(F)F</chem>	17				
	1H-Pyrazole-3-carboxylic acid, 1-[4-(aminosulfonyl)phenyl]-5-(4-chlorophenyl)-, methyl ester	<chem>COC(=O)c1cc(c2ccc(Cl)cc2)n(n1)c3ccc(cc3)S(=O)(=O)N</chem>	13				
	Benzenesulfonamide, 4-[3-(5-fluoro-4-methoxyphenyl)-3-(trifluoromethyl)-1H-pyrazol-1-yl]-	<chem>COc1ccc(cc1F)c2cc(m2c3ccc(cc3)S(=O)(=O)N)C(F)(F)F</chem>	11				
	Benzenesulfonamide, 4-[3-cyano-5-(4-methoxyphenyl)-1H-pyrazol-1-yl]-	<chem>COc1ccc(cc1)c2cc(m2c3ccc(cc3)S(=O)(=O)N)C#N</chem>	11				

Figure 6. Prediction output from the protocol built by Pipeline Pilot. The figure shows an example for WO1995015316(A1) (Celebrex). The leftmost column shows the structures of the top 5 predicted key compounds, followed by their chemical names, canonical smiles, the number of compounds similar to the predicted key compounds, and structures of similar compounds. It should be noted that the predicted key compound itself is included in the list of similar compounds.

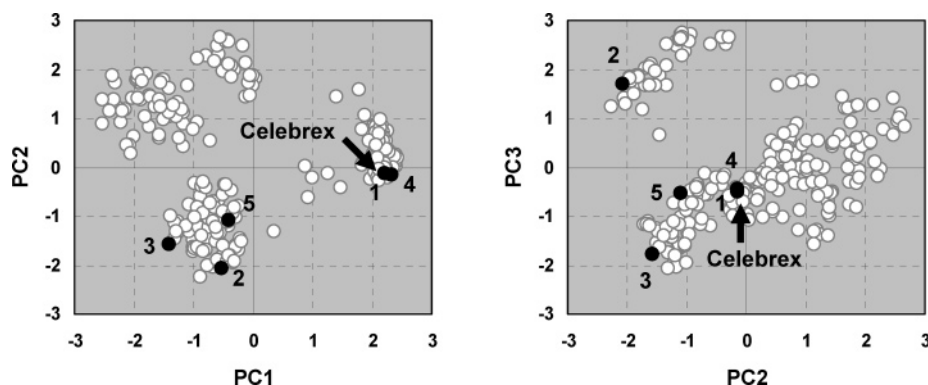


Figure 7. Predicted key compounds for WO1995015316(A1) shown on PCA plots. Black circles represent predicted key compound candidates, and numbers mean the rank order of them. Celebrex is indicated by arrows.

shown as an example. This would allow medicinal chemists to see not only the predicted key compounds but also their structural neighbors. Understanding representative structural classes within the example compounds would also be facilitated. Predicted key compounds were mapped on the PCA plots in Figure 7.

CONCLUSIONS

In this study, we have developed a computational method for predicting key compounds in a competitor's patent application. Traditionally, medicinal chemists make an educated guess on the identity of key compounds based on information such as biological data, scale of reaction, and/

or examination of the optimal salt form for a particular compound. However, this approach is inadequate when little information is available regarding key compounds.

In contrast, our method exclusively uses structural information on patent example compounds. Every chemical patent application states structural information explicitly in the examples section as chemical structures or chemical names; thus, in principle, our method is applicable to all patent applications. The validation study using 30 patents that contains launched drugs showed that our method was able to predict key compounds correctly in about 57% of the patents when ECFP_4 fingerprints and a similarity cutoff value of 0.7 were used as parameters.

One of the limitations of this method is it cannot be used immediately after a competitor's patent application is published, since it takes more than 1 month after the patent application is published for CA index names to be registered in the CAS databases and available, a requirement since we utilize Scifinder for the generation of a list of chemical names. In this kind of patent analysis, promptness is important as well as accuracy of prediction; thus, this issue should be addressed in future works, with a possible solution being the direct text mining of patent documents.

In addition to the utility described in this report, the method could be used to prepare the filing of patent applications from in-house drug discovery programs, i.e., the method would be used to check whether or not key compounds are easily revealed. If key compounds can be easily predicted, it indicates that there is a risk that competitors could identify our own key compounds with methods similar to the one presented in this study. In this case, it would be better to modify the distribution of example compounds so as not to allow easy identification of key compounds.

In summary, this is, to our knowledge, the first computational approach to predicting key compounds in competitors' patent applications. This method should prove useful for quality decision making in patent analysis.

ACKNOWLEDGMENT

The authors wish to thank Drs. Mamoru Uchiyama, Yoshiyuki Okumura, and Bruce A. Lefker for their fruitful discussions on the development of this work and Dr. Mitsuhiro Kawamura for critical reading of the manuscript.

REFERENCES AND NOTES

- Webber, P. M. Protecting your inventions: The patent system. *Nat. Rev. Drug Discovery* **2003**, *2*, 823–830.
- Atkinson, A. J.; Mellish, C.; Aitken, S. Combining information extraction with genetic algorithms for text mining. *IEEE Intell. Syst.* **2004**, *19*, 22–30.
- Davi, A.; Haughton, D.; Nasr, N.; Shah, G.; Skaletsky, M.; Spack, R. A review of two text-mining packages: SAS TextMining and WordStat. *Am. Statistician* **2005**, *59*, 89–103.
- Hale, R. Text mining: getting more value from literature resources. *Drug Discovery Today* **2005**, *10*, 377–379.
- Krallinger, M.; Valencia, A. Text-mining and information-retrieval services for molecular biology. *Genome Biol.* **2005**, *6*, 224.
- Erhardt, R. A. A.; Schneider, R.; Blaschke, C. Status of text-mining techniques applied to biomedical text. *Drug Discovery Today* **2006**, *11*, 315–325.
- Ananiadou, S.; Kell, D. B.; Tsujii, J. Text mining and its potential applications in systems biology. *Trends Biotechnol.* **2006**, *24*, 571–579.
- Cho, C. R.; Labow, M.; Reinhardt, M.; van Oostrum, J.; Peitsch, M. C. The application of systems biology to drug discovery. *Curr. Opin. Chem. Biol.* **2006**, *10*, 294–302.
- Cohen, A. M.; Hersh, W. R. A survey of current work in biomedical text mining. *Brief. Bioinform.* **2005**, *6*, 57–71.
- Hoffmann, R.; Krallinger, M.; Andres, E.; Tamames, J.; Blaschke, C.; Valencia, A. Text mining for metabolic pathways, signaling cascades, and protein networks. *Sci. STKE* **2005**, *2005* (283), 21.
- Jensen, L. J.; Saric, J.; Bork, P. Literature mining for the biologist: from information retrieval to biological discovery. *Nat. Rev. Genet.* **2006**, *7*, 119–129.
- Krallinger, M.; Erhardt, R. A.; Valencia, A. Text-mining approaches in molecular biology and biomedicine. *Drug Discovery Today* **2005**, *10*, 439–445.
- Larranaga, P.; Calvo, B.; Santana, R.; Bielza, C.; Galdiano, J.; Inza, I.; Lozano, J. A.; Armananzas, R.; Santafe, G.; Perez, A.; Robles, V. Machine learning in bioinformatics. *Brief. Bioinform.* **2006**, *7*, 86–112.
- Natarajan, J.; Berrar, D.; Hack, C. J.; Dubitzky, W. Knowledge discovery in biology and biotechnology texts: a review of techniques, evaluation strategies, and applications. *Crit. Rev. Biotechnol.* **2005**, *25*, 31–52.
- Rebholz-Schuhman, D.; Cameron, G.; Clark, D.; van Mulligen, E.; Coatrieux, J. L.; Del Hoyo Barbolla, E.; Martin-Sanchez, F.; Milanese, L.; Porro, I.; Beltrame, F.; Tollis, I.; Van der Lei, J. SYMBIOmatics: synergies in Medical Informatics and Bioinformatics—exploring current scientific literature for emerging topics. *BMC Bioinform.* **2007**, *8* (Suppl 1), S18.
- Roberts, P. M. Mining literature for systems biology. *Brief. Bioinform.* **2006**, *7*, 399–406.
- Scherf, M.; Epple, A.; Werner, T. The next generation of literature analysis: integration of genomic analysis into text mining. *Brief. Bioinform.* **2005**, *6*, 287–297.
- Schulze-Kremer, S. Ontologies for molecular biology and bioinformatics. *In Silico Biol.* **2002**, *2*, 179–193.
- Shatkay, H.; Feldman, R. Mining the biomedical literature in the genomic era: an overview. *J. Comput. Biol.* **2003**, *10*, 821–855.
- Spasic, I.; Ananiadou, S.; McNaught, J.; Kumar, A. Text mining and ontologies in biomedicine: making sense of raw text. *Brief. Bioinform.* **2005**, *6*, 239–251.
- Fattori, M.; Pedrazzi, G.; Turra, R. Text mining applied to patent mapping: a practical business case. *World Pat. Inf.* **2003**, *25*, 335–342.
- Dou, H. Benchmarking R&D and companies through patent analysis using free databases and special software: a tool to improve innovative thinking. *World Pat. Inf.* **2004**, *26*, 297–309.
- Grandjean, N.; Charpiot, B.; Pena, C. A.; Peitsch, M. C. Competitive intelligence and patent analysis in drug discovery. Mining the competitive knowledge bases and patents. *Drug Discovery Today: Technol.* **2005**, *2*, 211–215.
- Fischer, G.; Lalyre, N. Analysis and visualization with host-based software - The features of STN AnaVist. *World Pat. Inf.* **2006**, *28*, 312–318.
- Eldridge, J. Data visualization tools - a perspective from the pharmaceutical industry. *World Pat. Inf.* **2006**, *28*, 43–49.
- OmniViz. http://www.omniviz.com/applications/omni_viz.htm (accessed Sept 12, 2007).
- ClearForest. <http://www.clearforest.com/index.asp> (accessed Sept 12, 2007).
- Banville, D. L. Mining chemical structural information from the drug literature. *Drug Discovery Today* **2006**, *11*, 35–42.
- MDL Patent Chemistry Database. http://www.mdl.com/products/knowledge/patent_db/index.jsp (accessed Sept 12, 2007).
- JUBILANT'S Small Molecule Databases. <http://jubilantbiosys.com/smd.htm> (accessed Sept 12, 2007).
- Goodnow, R. A., Jr.; Gillespie, P. Hit and Lead identification: efficient practices for drug discovery. *Prog. Med. Chem.* **2007**, *45*, 1–61.
- Bleicher, K. H.; Böhm, H. J.; Müller, K.; Alanine, A. I. Hit and lead generation: Beyond high-throughput screening. *Nat. Rev. Drug Discovery* **2003**, *2*, 369–378.
- Goodnow, J. R. A. Hit and lead identification: Integrated technology-based approaches. *Drug Discovery Today: Technol.* **2006**, *3*, 367–375.
- Scifinder. <http://www.cas.org/products/scifinder/index.html> (accessed Sept 12, 2007).
- CambridgeSoft Batch Struct=Name. <http://www.cambridgesoft.com/software/details/?ds=5> (accessed Sept 12, 2007).
- Pipeline Pilot, version 5.0; Accelrys Inc.: San Diego, CA, 2005.
- Tanimoto coefficient, $T(a,b)$, is defined as $T(a,b) = (Na \& Nb) / (Na + Nb - Na \& Nb)$, where Na and Nb are the number of "1" bits in fingerprints a and b , respectively, and $Na \& Nb$ is the number of "1" bits appeared in both a and b fingerprints.
- MDL public keys; MDL Elsevier: San Ramon, CA. <http://www.mdl.com> (accessed Sept 12, 2007).
- Top 200 Drugs for 2005 by Sales. http://www.drugs.com/top200_2005.html (accessed Sept 12, 2007).
- Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–112.
- Dixon, S. L.; Koehler, R. T. The Hidden Component of Size in Two-Dimensional Fragment Descriptors: Side Effects on Sampling in Bioactive Libraries. *J. Med. Chem.* **1999**, *42*, 2887–2900.
- Flower, D. R. On the Properties of Bit String-Based Measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379–386.