Exploring Functional Group Transformations on CASREACT

Paul E. Blower, Jr.,* Glenn J. Myatt, and Michael W. Petras

Chemical Abstracts Service, 2540 Olentangy River Road, Columbus, Ohio 43202

Received June 27, 1996[⊗]

CAS analyzed all single step reactions on the CASREACT reaction database in terms of functional group transformations. This involved defining over 1300 functional groups and a set of rules to determine exactly which functional groups participate in a particular reaction. This paper describes the analysis process and presents some statistics and charts showing the frequency distribution of functional groups and their transformations over the full database.

1. INTRODUCTION

Over the last decade a number of chemical reaction databases have been made available through structure-searchable computer systems including CASREACT on STN,[1,2] REACCS,[3] ORAC,[4] and Crossfire.[5] Although the quality and content of reaction data are important factors, the methods provided to search the databases also determine the usefulness of such systems. The traditional approach to searching such databases has been to perform a reaction substructure search. An example search is seen in Figure 1. Reactant and product structural fragments are defined, in addition to marking the reaction site(s) and assigning mappings between the reactants and products. Having specified the query, a substructure search is performed to return an answer set, as seen in Figure 1. This approach to searching reaction databases is very powerful when the question can be easily described as a substructure query. It is an extremely flexible method allowing any reactant and/or product substructure to be searched. Unfortunately, not all questions one would like to ask of a reaction database can be easily answered using this approach, and searching can be quite slow for small, common query structures. An effective scheme for classifying or indexing reactions would enhance the usability of the database by providing a more organized and direct route to reactions.

A number of authors have reported methods of classifying reactions.[6−17] Rohde has performed an analysis over the Theilheimer database for a number of reasons, including the automatic generation of transforms for the CASP program.[6] The method is based on the definition of a reaction type which defines the reacting atoms and bonds along with significant adjacent bonds. Along similar lines, Hendrickson developed a method of classifying reaction databases based on the "net structural change" or those bonds made and broken.[7−9] One application of the analysis was to provide literature precedents for schemes identified by the SYNGEN synthesis design program. Additionally, this classification scheme has been incorporated within the COGNOS program as a way of searching for reactions. Another approach adopted by Blurock makes use of reaction patterns, defined as two sets of automatically derived substructures.[10−12] These patterns include enough information to adequately describe a reaction center. The primary purpose of the analysis was for use in the synthesis planning program RETROSYN. Finally, the HORACE program hierarchically classifies sets of reactions based on both topological (which includes functional groups) and also physicochemical features.[14,15]

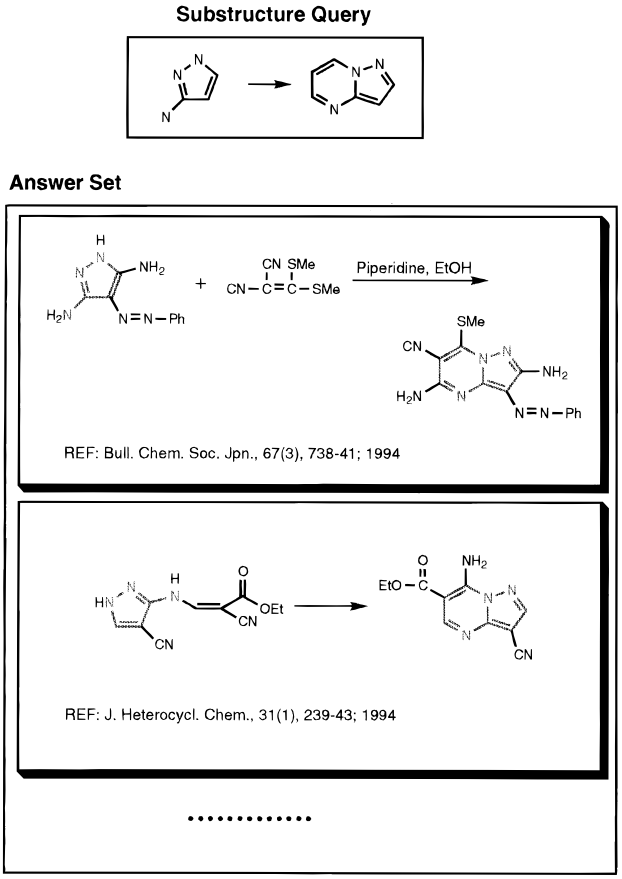


Figure 1. A traditional substructure query against a reaction databases.

The approach adopted at Chemical Abstracts Service (CAS) to classify the CASREACT file is by functional group transformations, i.e., the formation of a product functional group from one or more reacting functional groups. This paper describes how the analysis was performed, presents statistics showing the frequency distribution of functional groups and functional group transformations, and discusses potential uses of the analysis tables.

[⊗] Abstract published in *Advance ACS Abstracts,* December 15, 1996.

S0095-2338(96)00089-3 CCC: $14.00

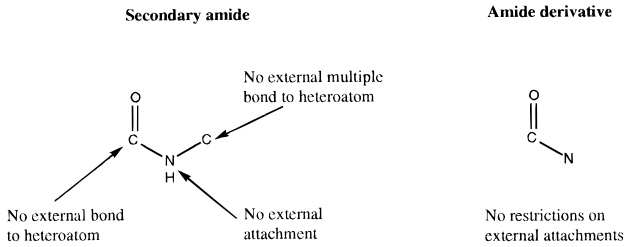


Figure 2. Functional groups definitions.

2. METHODOLOGY

2.1. Overview. This section gives an overview of the CASREACT file and presents details on how functional groups are identified, along with what types of functional groups are found. In addition, the rules required to determine those functional groups involved in a reaction are shown. This section closes with a description of the method used to generate a relational database summarizing the analysis.

2.2. The CASREACT File. CASREACT is a document-based, structure-searchable file containing chemical reaction information from documents covered in the Organic Sections of Chemical Abstracts. Journal coverage is from 1985 to the present and patents from 1990 to the present. In June 1996, the file contained more than 149 000 documents, covering approximately 1 300 000 single-step reactions.

CASREACT records contain reaction diagrams, CAS Registry Numbers for all reactants, products, reagents, solvents, and catalysts, yields for products, and textual reaction information. The reactants, reagents, and products are structure-searchable as are reaction sites and the atom map between reactant and product.

Since 1993, common functional groups found in reactants, reagents, and products have been searchable. With functional group terms, a user can ask specifically for the reactions or formation of a functional group or for nonreacting groups. This makes it possible to find very precise answers to questions involving functional group transformations, e.g., conversion of an acyclic ketone to a secondary alcohol.

2.3. Functional Groups. 2.3.1 Defining Functional Groups. Functional groups are named substructures defined as query connection tables and identified using common structure search techniques.[18] These connections tables often include generic atoms and/or bonds, such as M for any metal. In addition to defining the essential structure of the functional group, further restrictions are placed on the atoms and bonds in the external environment of the group. Figure 2 illustrates this for a *secondary amide* where a number of restrictions are placed on the atoms. For example, there can be no further heteroatom attachments to the carbonyl carbon. Where two functional groups covering the same set of atoms and bonds are identified, the smaller functional group is usually discarded. In the case of the two groups shown in Figure 2, whenever a *secondary amide* is identified the *amide derivative* is also found. Since the purpose of defining *amide derivative* is to collect only those amide-like substructures that do not satisfy the more precise amide terms, the *amide derivative* term would be dropped.

2.3.2. Basic Functional Groups. An extensive set of 255 traditional functional groups has been defined, using the format described in the previous section. These include functional groups such as ketone, aldehyde, carbamate, etc. This set contains precise definitions for groups containing the elements C, N, O, S, and the halogens. This collection of functional groups also has some less precise groups involving the elements Al, B, P, Se, Si, and Sn. There are a number of metal groups, but these are not covered at the same level of detail.

Table 1. Base Function Groups within Composite Groups

acid	amidine	cyclopropyl	iodide, alkyl
alcohol, alkyl	amine, alkyl	enamine	iodide, aryl
alcohol, aryl	amine, aryl	enol	ketone
alcohol, deriv	amine, deriv	epoxide	nitrile
aldehyde	bromide, alkyl	ester	thiocarboxylate
alkene	bromide, aryl	fluoride, alkyl	thiol, alkyl
alkyne	chloride, alkyl	fluoride, aryl	thiol, aryl
amide	chloride, aryl	imine	thiol, deriv

2.3.3. Heterocyclic Groups. The definition of functional groups has been extended to include a set of 194 heterocycles, such as *pyrrole* and *imidazole*. This extension reflects the importance of reactions forming heterocycles, particularly to the pharmaceutical industry. There are 73 rings with ring sizes of three to seven members and containing the elements C, N, O, and S. Once the rings have been identified at an atom level, they are categorized according to their level of unsaturation: aromatic (e.g., *imidazole*), fully saturated (e.g., *imidazolidine*), and partially saturated. In addition three common two-ring systems are defined: *penam*, *cephem*, *purine*.

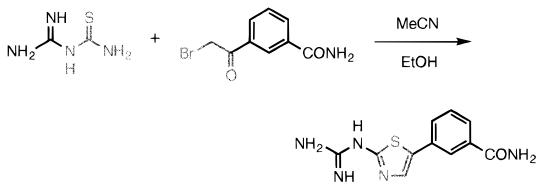
2.3.4. Composite Functional Groups. Composite functional groups are pairs of common groups that are in close proximity, separated by one or two bonds; e.g., *β-ketoester* and α-*aminoacid*. Composite groups were defined for two reasons: the component groups often react in concert and their proximity can strongly influence reactivity. Unlike the functional groups described above, the composite groups are not defined as explicit substructures. Instead they are generated by the program from the set of base groups listed in Table 1.

2.3.5. Unfunctionalized Substructures. Many common reactions, such as the alkylation of an active methylene, do not involve functional groups *directly*. In order to classify the entire CASREACT database, we added a series of unfunctionalized substructures to extend the scope of the analysis. In effect, this extends the concept of a functional group to include any named substructure. Examples of named substructures used in this analysis include *active methylene*, *C−C acyclic*, *and cyclohexyl*. To avoid generating spurious nonreacting functional groups, these named substructures are only recognized when participating in a reaction.

2.4. Functional Group Transformation and Rules. Each reaction is analyzed as a series of functional group transformations which are encoded as one or more n-tuples comprised of

- one formed product group
- 1−4 reactant groups that are transformed to the product group
- a code indicating how the reactant groups are distributed among the reacting substances.

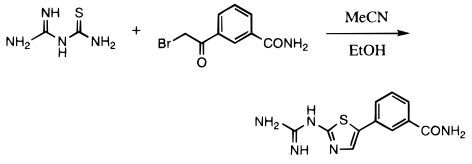
Figure 3 illustrates this process. Initially functional groups are identified in reactants and products, including α-*bromoketone*, *thiourea*, and *thiazole* which are examples of a composite group, a traditional group, and a heterocycle,



Representation:

Reactant	Reactant	Product
thiourea	α-bromoketone	thiazole

Figure 3. Analysis of a reaction by functional group transformation.



Product	Reactant	Reactant	Reactant	Rxns	Docs	Yield
thiazole	**thiourea**	*α-bromoketone*		295	62	72
thiazole	**thiourea**	*acyclic ketone*	*1°-alkyl bromide*	176	47	73

Figure 4. Multiple analysis lines from a single reaction.

respectively. The process for generating transformations works from the product group formed in the reaction, in this case the *thiazole*. A series of rules is used to determine exactly which reactant functional groups contribute to the formation of this group. The basic rule states that a reactant group (Ri) participates in the transformation if either

1. Ri is transformed to the product functional group, or
2. Ri is combined with Rj which is transformed to the product functional group.

There are additional rules to deal with special cases like partially overlapping reactant groups. All reactant groups that satisfy the rules are included in the transformation.

2.5. Producing the Analysis Tables. For each single step reaction that has a yield of over 25%, the analysis as described in section 2.4 is performed. It should be noted that each reaction may give rise to more than one table entry as illustrated in Figure 4. Both table entries generated are good descriptions for this reaction. Although the two entries are somewhat redundant, each provides important additional details. One entry describes the position of the bromide relative to the ketone and the other provides more detail of the functional group's environment, i.e., the bromide is primary alkyl and the ketone is acyclic.

The result of this analysis over the full CASREACT file is a large table of transformation entries. Each line in the table summaries all occurrences of a particular transformation; i.e., a unique combination of product and reactant functional groups and the grouping code. Three additional fields are recorded for each table entry: reaction and document frequencies, and the average yield. The number of documents is a useful indication of the scope of the reactions, since it is possible that a functional group transformation could appear in many reactions but only in a few documents. The average yield over all the reactions for a particular transformation gives evidence of the utility of the reaction class. Figure 4 details two lines that are taken from the final table. The different fonts are used to indicate how the functional groups are distributed over the reacting substances. For example, on the second line, the acyclic ketone and the primary alkyl bromide are in the same substance.

An additional relational table is constructed that is a valuable byproduct of the analysis performed. For each reaction, those functional groups that are not reacting are stored. This analysis is useful for studying the scope of a reaction in terms of known groups that can be carried through the transformation unaltered.

The analysis as described is extremely computer intensive. Even taking advantage of the massively parallel computing environment at CAS, the analysis still took over a week to complete.

3. RESULTS

3.1. Functional Group Statistics. The table shown in Table 2 indicates the high frequency functional groups that have been found in products of reactions on CASREACT. Not surprisingly benzene is contained in nearly three quarters of a million products. It can be seen from the table that the frequency of the top 14 functional groups for each class (traditional, heterocycles, and composite) drops quite rapidly. This distribution can be seen for all functional groups in the pie chart in Figure 5. The chart is divided into six nonequal segments. The chart shows, for example, that there are 147 functional groups with counts of 1−10. The chart segments increase logarithmically and apart from the two extreme groups 1−10 and 10 000+, the size of each segment is very similar.

3.2. Transformation Statistics. The table of unique functional group transformations contains approximately 1.5

Table 2. High Frequency Product Groups

traditional groups		heterocycles		composite groups	
freq	group	freq	group	freq	group
743 244	benzene	62 659	tetrahydrofuran	84 304	1,2-diene
251 245	carboxylate	60 912	tetrahydropyran	60 154	α,β-unsatd-ketone
246 171	ether	595 43	pyridine	47 598	α,β-unsatd-ester
122 728	cyclic olefin	44 730	pyrrole	31 257	α-d_amino-amide[19]
108 275	cyclic ketone	38 057	pyrrolidine	30 472	α-d_amino-ester[19]
104 248	acyclic olefin	37 611	1,3-dioxolane	26 914	1,2-di-alkyl alcohol
103 042	acetyl	37 509	pyridine-H	25 042	allyl alcohol deriv
95 607	acyclic ketone	32 493	pyran-H	24 080	α,β-unsatd-amide
91 220	C-metal	30 968	imidazole	23 688	1,2-diol deriv
80 064	terminal olefin	30 459	piperidine	20 424	1,2-diol mono deriv
80 033	silyl	29 201	pyrrole-H	19 767	1,3-diene
76 103	acetal	29 096	pyrimidine-H	17 964	1,3-hydroxy-alkene
72 338	2° amide	26 825	pyrimidine	17 060	1,2-hydroxy-amine deriv
69 816	1° acyclic alcohol	18 787	oxirane	16 377	1,2-alkenyl-amine deriv

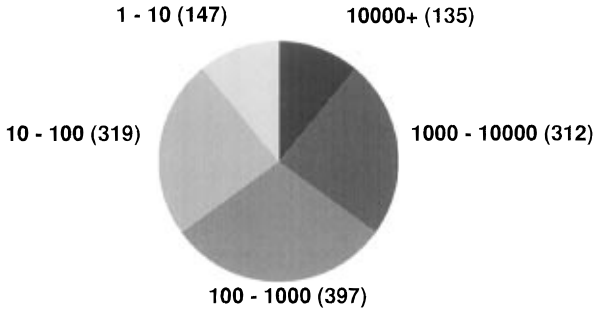


Figure 5. Distribution of product functional groups.

Table 3. High Frequency Transformations

freq	product	reactant	reactant
19726	carboxylic acid	carboxylate	
12282	primary acyclic alcohol	carboxylate	
10166	C−H unfunctionalized	cyclic olefin	
9039	C−H unfunctionalized	terminal olefin	
9005	secondary cyclic alcohol	carboxylate	
8378	secondary amide	carboxylate acid	primary amine
8037	secondary acyclic alcohol	acyclic ketone	
7279	primary amine	carbamate	
7161	secondary cyclic alcohol	cyclic ketone	
6867	C−H unfunctionalized	acyclic olefin	
6382	acyclic olefin	active methylene	aldehyde
5677	phenol	ether	
5477	C−H unfunctionalized	cyclohexene	
5380	carboxylate	carboxylic acid	

million lines. Table 3 displays the high frequency transformations. There are 19 726 reactions on the CASREACT file indexed by the carboxylate to carboxylic acid transformation and 12 282 by the carboxylate to primary acyclic alcohol transformation. The graph in Figure 6 is a log−log plot of the frequency distribution for the entire table. Each point on the graph is the number of table entries that have a particular reaction frequency. For example, the point on the far right is the first entry in Table 3, i.e., there is one functional group transformation that indexes 19 726 reactions. The point plotted on the far left indicates that there are 845 390 entries in the table with only one reaction.

Table 4. Comparison of Structure-Based and Transformation Index Searching

substructure search	functional group transformations
can result in time consuming searches	extremely fast, relying on a precomputed index
answer set determined at run-time	precomputed answer sets
cannot browse reaction classes	browse reaction classes
can specify complex queries	can only search functional groups that have been predefined
very flexible	precomputed index
can only answer specific questions	can be used as discovery/learning tool

4. DISCUSSION

The analysis table of unique functional group transformations provides new opportunities to enhance the usefulness of the CASREACT database. Each table entry is essentially an index to the reactions involving the transformation. The analysis table is also a highly descriptive summary of reaction classes on the database.

Used as an index of functional group transformations, it offers an alternative and complementary approach to traditional substructure-based reaction searching. The chemist could designate a portion of the index for viewing by placing restrictions on the index entries displayed, then select index entries, and browse the corresponding reactions. For example, one could ask to see all reactions that form hydrazines in greater than 50% yield and involving only one reactant. Figure 7 illustrates this process. In contrast to a structure-based search which can be quite slow, this type of searching is extremely fast, since it is simply a direct look-up of a precomputed index. Table 4 summarizes the two approaches to searching reaction databases.

The analysis table is also useful as a high level summary that gives a complete picture of the database content. Figure

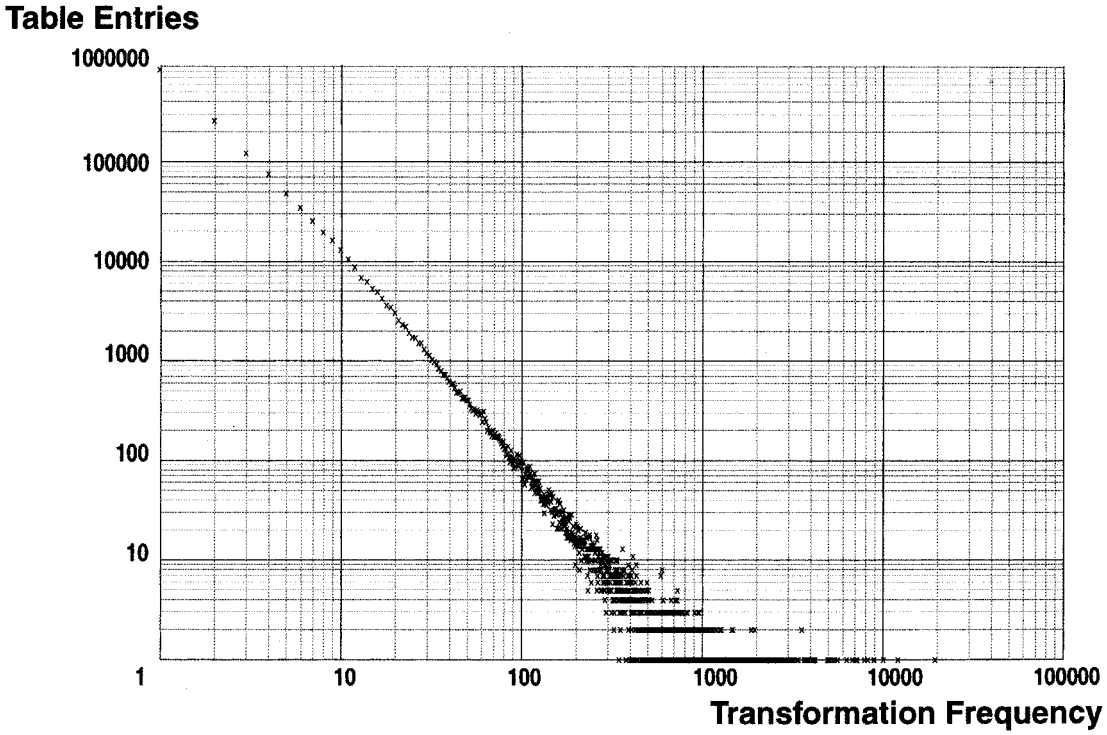


Figure 6. Transformation statistics.

Product	Reactant	Rxns	Yield
hydrazine	hydrazone	138	74
hydrazine	azo	85	67
hydrazine	1°-amine	85	67
hydrazine	amino benzene	61	72
hydrazine	amidine drv	57	64
hydrazine	hydrazide	50	66
hydrazine	pyrazole	45	65
hydrazine	2°-amine	44	74
hydrazine	nitrosamine	42	76

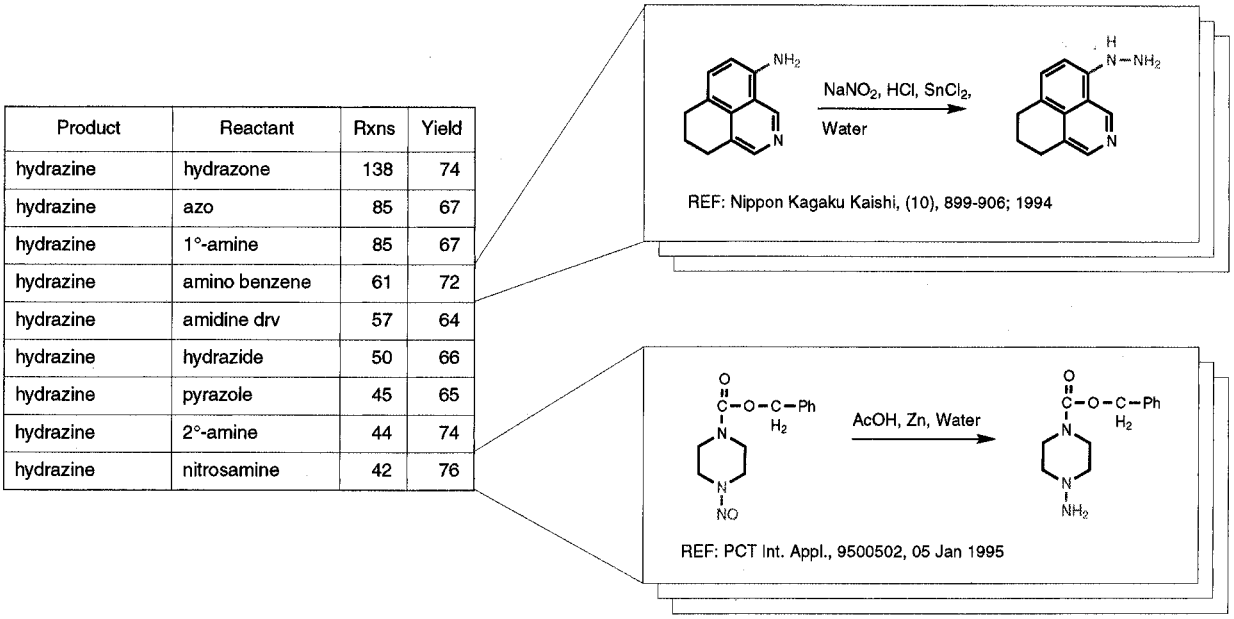


Figure 7. Searching reaction databases via a functional group transformation index.

7 shows a portion of the table where all reactions involving the formation of a hydrazine group are clustered together, with statistics for each cluster. Browsing this portion of the index gives a global view of all hydrazine forming reactions on the database, clustered by reactant groups.

An effective reaction classification scheme could also have practical applications in other areas including the comparison of reaction databases, assessing the novelty of new reactions to be added to the database, and the automatic generation of transforms for use in synthesis design knowledge bases. A reaction summary of this kind could be used as a learning/discovery tool.

5. CONCLUSION

Functional groups are a precise, well-established way of describing the important components involved in a reaction. We analyzed all reactions in the CASREACT database in terms of functional group transformations. This analysis is both an exhaustive index to the database and a highly descriptive summary of the reaction classes reported in the literature over the last 11 years.

As an index, the analysis offers an alternative and complementary approach to traditional substructure-based reaction searching. In addition to asking very specific queries using the traditional approach, the index of functional group transformations allows the chemist to ask much broader questions.

The analysis table also provides a high level summary that gives a complete picture of the database content. This can be equally useful as a learning tool, helping the chemist to discover unexpected relationships.

ACKNOWLEDGMENT

We would like to thank Robert E. Stobaugh and Dannie J. Saunders for their assistance in this project.

REFERENCES AND NOTES

(1) Blake, J. E.; Dana, R. C. CASREACT: More than a Million Reactions. *J. Chem. Inf. Comput. Sci*. **1990**, *30*, 394−399.

(2) CASREACT on STN and SciFinder is available from Chemical Abstracts Service, Columbus, OH.

(3) REACCS is available from MDL Information Systems, Inc., San Leandro, CA.

(4) Johnson, A. P.; Cook, A. P. In *Modern approaches to chemical reaction searching*; Willett, P., Ed.; Gower: Aldershot, 1986; pp 184−201.

(5) Crossfire plus reactions is available from Beilstein Informationssysteme GmbH, Frankfurt, Germany.

(6) Rohde, B. Reaction type informetrics of chemical reaction databases: how "large" is chemistry? *Spec. Publ. - R. Soc. Chem*. **1994**, *142* (Further Advances in Chemical Information), 109−127.

(7) Hendrickson, J. B., Miller, T. M. Reaction Classification and Retrieval. A Linkage between Synthesis Generation and Reaction Databases. *J. Am. Chem. Soc*. **1991**, *113*, 902−910.

(8) Hendrickson, J. B.; Sander, T. COGNOS: A Beilstein-Type System for Organizing Organic Reaction. *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 251−260.

(9) Hendrickson, J. B.; Miller, T. M. Reaction Indexing for Reaction Databases, *J. Chem. Inf. Comput. Sci*. **1990**, *30*, 403−408.

(10) Blurock, E. S. Reaction: System for Modeling Chemical Reactions. *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 607−616.

(11) Blurock, E. S. Computer-Aided Synthesis Design at RISC-Linz: Automatic Extraction and Use of Reaction Classes. *J. Chem. Inf. Comput. Sci*. **1990**, *30*, 505−510.

(12) Blurock, E. S. Automatic Extraction of Reaction Information from Databases Using Classification and Learning Techniques. *Chem. Inf*. **1990**, *2*, 25−35.

(13) Gelernter, H.; Rose, J. R.; Chen, C. Building and Refining a Knowledge Base for Synthetic Organic Chemistry via the Methodology of Inductive and Deductive Learning. *J. Chem. Inf. Comput. Sci*. **1990**, *30*, 492−504.

(14) Rose, J. R.; Gasteiger, J. HORACE: An Automatic System for the Hierarchical Classification of Chemical Reactions. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 74−90.

(15) Chen, L.; Gasteiger, J.; Rose, J. R. Automatic Extraction of Chemical Knowledge from Organic Reaction Data: Addition of Carbon-Hydrogen Bonds to Carbon−Carbon Double Bonds. *J. Org. Chem*. **1995**, *60*, 8002−8014.

(16) Fujita, S. "Structure-Reaction Type" Paradigm in the Conventional Methods of Describing Organic Reactions and the Concept of Imaginary Transition Structures Overcoming This Paradigm. *J. Chem. Inf. Comput. Sci*. **1987**, 27, 120−126.

(17) Weise, A. Synthesis Simulation by Synthon Substitution. *J. Chem. Inf. Comput. Sci*. **1990**, *30*, 490−491.

(18) Stobaugh, R. E. Chemical Structure Searching. *J. Chem. Inf. Comput. Sci*. **1985**, *25*, 271−275.

(19) d_amino is an amine derivative.

CI960089L