0.221 mmol) in 3 mL of $CFCl_3$ at -78 °C in a Rayonet photoreactor produced the bicyclopentane **2b** as shown by its low temperature NMR.[18] When nitroxide (50.0 mg, 0.263 mmol) was added to this solution and the reaction allowed to warm up to room temperature, the same amounts of the adducts *trans*-**3b** and **3b'** were produced as in the direct photolysis of azoalkane **1b**.

Acknowledgment. Financial support by the Deutsche Forschungsgemeinschaft and the Fonds der Chemischen Industrie are much appreciated. S.E.B. thanks the Alexander von Humboldt Foundation for the generous provision of a postdoctoral fellowship (1990/91).

Registry No. **1a**, 2721-32-6; **1b**, 31689-32-4; **1c**, 66322-90-5; **1d**, 138062-33-6; **2a**, 185-94-4; **2b**, 72447-89-3; *cis*-**3a**, 134278-19-6; *trans*-**3a**, 134278-20-9; *cis*-**3b**, 138062-37-0; *trans*-**3b**, 138062-39-2; **3b'**, 138062-38-1; *cis*-**3c**, 138062-41-6; *trans*-**3c**, 138062-40-5; **4d**, 138062-42-7; 1,4-dibenzoylbenzene, 3016-97-5; cyclopentadiene, 542-92-7; 1,4-bis[(2,4-cyclopentadien-1-ylidenephenyl)methyl]-benzene, 138062-34-7; 1,4-bis[2,3-bis(ethoxycarbonyl)-7-(phenylmethylidene)-2,3-diazabicyclo[2.2.1]hept-5-en-7-yl]benzene, 138062-35-8; 1,4-bis[2,3-bis(ethoxycarbonyl)-7-(phenylmethylidene)-2,3-diazabicyclo[2.2.1]heptan-7-yl]benzene, 138062-36-9; diethyl azodicarboxylate, 1972-28-7; 2,3-diazabicyclo[2.2.2]oct-2-ene, 3310-62-1; 1,1,3,3-tetramethyl-1,3-dihydroisoindolin-2-yloxyl, 80037-90-7.

Supplementary Material Available: X-ray crystallographic data for *trans*-**3a** and *trans*-**3b**, including atomic coordinates, equivalent isotropic displacement parameters, bond lengths, bond angles, anisotropic displacement parameters, H atom coordinates, and isotropic displacement parameters, and ^{13}C and ^{1}H NMR spectra for **1d** and its precursors, namely the diethyl azodicarboxylate adduct and its hydrogenated derivative (12 pages). This material is contained in many libraries on microfiche, immediately follows this article in the microfilm version of the journal, and can be ordered from the ACS; see any current masthead page for ordering information.

Reaction Retrieval from Databases for Organic Chemists

James B. Hendrickson* and Todd M. Miller

Edison Chemical Laboratories, Brandeis University, Waltham, Massachusetts 02254-9110

Received May 29, 1991

A general rationalization of organic reactions is described, based on the net structural change at the skeletal atoms in the reacting center. We define four generalized kinds of attachments any skeletal atom may have and how they change in any reaction. Unit reactions are defined as unit exchanges of these attachments. With this logical basis a simple organization of all organic reactions is developed to provide a logical overview of all organic reactions. A program, RETRIEVE, is described to index all the entries in two common reaction databases and to retrieve all precedents for a given reaction. A statistical breakdown of the reaction types in these databases is presented. Most entries are just single or double unit reactions and can be quickly accessed and closely matched to the reactive center of any input query.

In organic chemistry there is a long history of searching catalogs for structures, but very little organized indexing for reactions. Indexing name reactions is common but these are haphazard, gratuituous, and not systematic. Now that computer databases of reactions have made an appearance there is a need for a logical basis for indexing organic reactions. This requires a general system to describe the net structural change in any reaction in terms of what bonds are made and broken. Such a system should proceed from the general to the particular to allow for nesting of categories, and it must provide that every possible reaction change have a place in the indexing format, regardless of whether there are known examples. This idea reflects the successful Beilstein system for indexing structures in that any structure, known or not, has a place in the system. This assures that if any compound is included it will be found and, if not included, one is assured that it is unknown.

In this paper we present such an indexing system for reactions and describe a program, called RETRIEVE, which uses it for the retrieval of reactions for computer databases. We will show that most reactions in two large popular databases can be efficiently indexed and reliably retrieved by the use of this simple system, which also provides a useful understanding of their contents. In turn, the successful application of the system confirms the validity of the system itself for organizing and cataloguing organic reactions.

To describe all organic reactions we need to define a general, abstract definition of the nature of bond changes to accommodate all possible instances, independent of mechanism. It is important to focus only on the *net structural change* at the reacting centers themselves. The remaining atoms, which do not change, are not relevant to the search. Hence, a search procedure organized by structural similarities at *all* the atoms is likely to miss good precedents in which the reacting centers are the same but the remainder of the molecule may be quite different.

System of Reaction Description. The system we apply is based on a structural premise of a backbone skeleton of linked carbons which bear functionality in the form of π-bonds and attached heteroatoms.[1] Two kinds of heteroatom attachments are distinguished initially, electropositive and electronegative, to afford recognition of the oxidation state of the carbon to which they are attached. Isohypsic attachments[2,3] to another carbon are distinguished between the σ-bond, the skeletal attachment, and the π-bond, a functional group. This creates a fundamental and generalized definition of four synthetically important kinds of attachments at any carbon in a structure: H for a bond to hydrogen or electropositive atom (B, Al, Si, Sn, metals); R for a σ-bond to carbon (skeletal); Π for a π-bond to carbon (functional); Z for any bond (π- or σ-) to an electronegative heteroatom (N, O, S, Hal, etc.)

(1) This backbone skeleton may incorporate atoms other than carbon (cf., N, O, S), but it is simpler first to define in terms of a carbon skeleton, leaving the others for discussion later; they will be seen to follow the same procedure exactly.

(2) The term *isohypsic*, from the Greek for "equal level", was introduced to mean neither oxidative nor reductive.[3]

(3) Hendrickson, J. B. *J. Am. Chem. Soc.* **1971**, *93*, 6847.

0022-3263/92/1957-0988$03.00/0

Table I. 16 Possible Unit Exchanges at Any Skeletal Carbon

substitution	HH, ZZ, RR, ΠΠ	Δx 0	$\Delta\pi$ 0	$\Delta\sigma$ 0
oxidation	ZH	+2	0	0
reduction	HZ	−2	0	0
elimination	ΠH	+1	+1	0
	ΠZ	−1	+1	0
addition	HΠ	−1	−1	0
	ZΠ	+1	−1	0
construction	RH	+1	0	+1
	RZ	−1	0	+1
	RΠ	0	−1	+1
fragmentation	HR	−1	0	−1
	ZR	+1	0	−1
	ΠR	0	+1	−1

The *number* of bonds of each kind is defined as h, σ, π, z, respectively, with a sum of 4. Hence there are only three variables needed to describe the state of any carbon, the fourth obtained by difference, analogous to familiar structure drawings in which the number of hydrogens is recognized by difference. Since the oxidation state of any carbon is simply given by $x = z-h$, the three variables for any carbon may be taken as x, π, σ. This system sharply distinguishes between the functionality of a molecule and its backbone skeleton, expressed as skeletal attachments R, with σ as the skeletal valence, i.e., $\sigma = 1$, primary; $\sigma = 2$, secondary, etc. The functionality of the carbon can then be represented by just two digits,[4] z and π.

For a reaction the *net structural change* at any carbon is also expressed by three variables: Δx, $\Delta\pi$, $\Delta\sigma$. These describe familiar categories of reactions, i.e., oxidation/reduction, elimination/addition, construction/fragmentation, respectively, as three pairs with one change in each the reverse of the other. A unit exchange of attachments at any carbon may be described with two letters, the first for the bond formed, the second for the bond broken. For the four kinds of attachments there are therefore 16 possible unit exchanges, summarized in Table I with their characteristic changes in the three variables.

More than one changing carbon may be involved in a reaction; these will be skeletally adjacent and coupled. A unit exchange of ±H or ±Z requires no other carbon, but one which contains ±R or ±Π must involve an adjacent, coupled carbon also changing with the same ±R or ±Π. Thus a reductive addition to a π-bond involves two carbons, as HΠ.HΠ, and a simple construction is RH.RZ, changing at two carbons. Taken together these coupled unit exchanges constitute a *unit reaction*.

Refunctionalization reactions are defined as those which involve no skeletal change ($\Delta\sigma = 0$) and so include all unit exchanges without ±R. These unit reactions almost never involve more than three adjacent carbons changing; in the databases examined 99% of the refunctionalizations involve no more than three carbons.[5] Therefore, virtually all unit refunctionalizations can be gathered into a kind of "periodic table" of 14 unit reactions, illustrated in Table II and organized by the two remaining variables: oxidation state change, $\sum\Delta x$, and π-bond change, $\sum\Delta\pi$, summed over three coupled skeletal carbons. In each group in Table

(4) One could as well use x and π but the use of z and π is a more direct description.

(5) The next vinylogy after the substitutions would be that of the additions/eliminations, e.g., 1,4-addition of Br_2 to a diene, ZΠ.ΠΠ.ΠΠ.ZΠ on four coupled carbons. These are not included in Table II owing to their rarity in the present databases but such an extension is a simple, logical one.

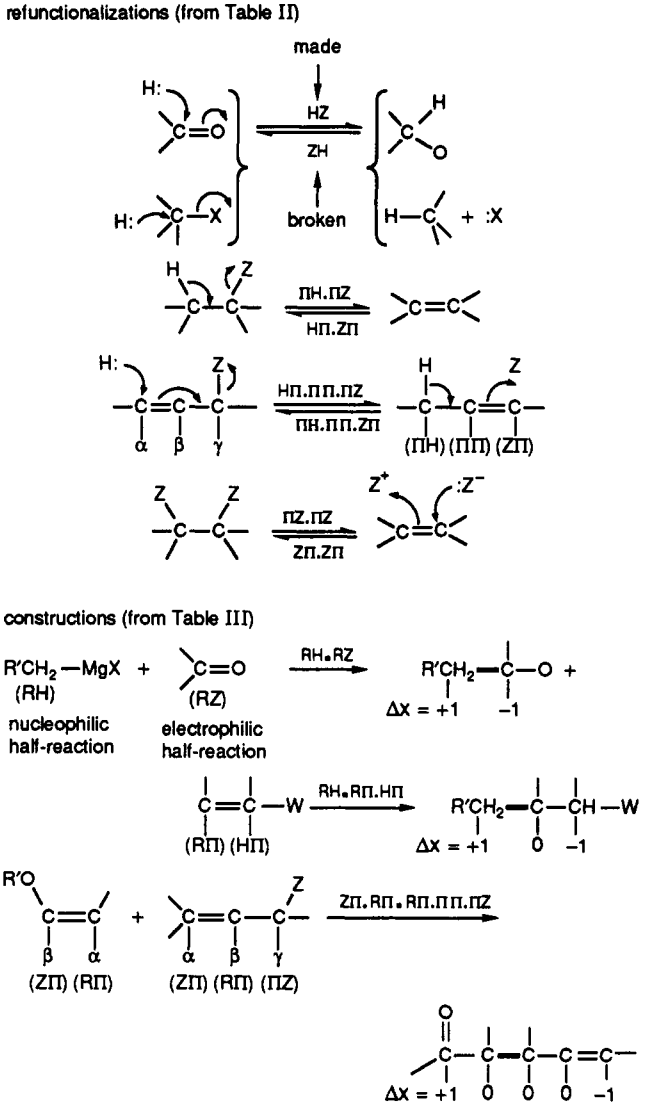

Figure 1. Unit reaction types illustrated.

II the box shows both the unit reaction in terms of the attachments (H, R, Π, Z) made and broken and also a simple designation of the group in brackets, in terms of the type (at right) and oxidation state change (at bottom). Simple substitutions of H and Z require only one changing carbon, additions and eliminations have two coupled carbons, and three coupled carbons are characteristic of the allylic, or vinylogous, substitutions.[5] The simple isohypsic substitutions, HH and ZZ, are designated as [H] and [S], respectively, and their allylic counterparts as [H'] and [S']. Several reaction designations are illustrated with their traditional structure drawings in Figure 1.

Actual reactions of course may involve more than one unit reaction, as successive steps. Among refunctionalizations the common cases of two unit reactions together are the same reaction at two sites simultaneously (cf., reduction of two different ketones, HZ) or two successive ones at the same site (cf., –COOH → $-CH_2OH$, which is HZ + HZ). However, our survey of the databases described below shows that some 78% of the refunctionalizations are in fact single unit reactions and 12% consist of only two unit reactions; thus, fully 90% of these reactions are simply described in this way.

Constructions and their reverse, fragmentations, alter the skeleton and so involve the third reaction variable, i.e., $\Delta\sigma$. These reactions may be seen as composed of two unit *half-reactions* which join together at each ±R skeletal

Table II. Unit Refunctionalizations on 1-3 Skeletal Atoms

$\Sigma\Delta x =$	-2	0	+2		
$\Sigma\Delta\pi =$	[RE]	[E]	[XE]		
+2	ΠZ.ΠZ	ΠH.ΠZ	ΠH.ΠH	Elimination [E]	(2 atoms)
	[R]	[H], [S]	[X]		
0	HZ	HH , ZZ	ZH	Simple Substitution [S]	(1 atom)
	[R']	[H'], [S']	[X']		
0	HΠ.ΠΠ.ΠZ	HΠ.ΠΠ.ΠH ZΠ.ΠΠ.ΠZ	ZΠ.ΠΠ.ΠH	Allylic Substitution [S']	(3 atoms)
	[RA]	[A]	[XA]		
-2	HΠ.HΠ	HΠ.ZΠ	ZΠ.ZΠ	Addition [A]	(2 atoms)
	Reductive	Isohypsic	Oxidative		
	[R]	—	[X]		

Table III. Construction/Fragmentation Half-Reactions on 1-3 Skeletal Atoms

	Reductive		Oxidative			
$\Sigma\Delta x =$	-1	-1	+1	+1		
$\Sigma\Delta\pi = +2$	—	ΠR.ΠZ	ΠR.ΠH	—	Elimination	(2 atoms)
0	RZ	HR	ZR	RH	Simple Substitution	(1 atom)
0	RΠ.ΠΠ.ΠZ	HΠ.ΠΠ.ΠR	ZΠ.ΠΠ.ΠR	RΠ.ΠΠ.ΠH	Allylic Substitution	(3 atoms)
-2	RΠ.HΠ	—	—	RΠ.ZΠ	Addition	(2 atoms)
	Construction	Fragmentation	Fragmentation	Construction		
	(Electrophilic)			(Nucleophilic)		

carbon, the two carbons which are being linked or fragmented. Like the refunctionalizations each half-reaction is also commonly limited to three adjacent changing carbons each, seen as linear strands of carbons out from each constructing (or fragmenting) carbon. The changing atoms in these strands are labeled, in order out from the ±R carbon, as α, β, γ. Half-reactions which change at one atom only must also incorporate only ±H or ±Z and only change the α atom, i.e., RH, RZ, HR, ZR. Additions/eliminations involve ±Π and so must occur at two atoms, α and β, and those at three atoms are allylic, at α, β, γ. An analogous "periodic table" of all possible unit half-reactions, on strands of three carbons, is shown in Table III. The nucleophilic half-reactions in construction are oxidative ($\sum\Delta x = +1$), the electrophilic ones are reductive ($\sum\Delta x = -1$). Almost all full constructions are simple unit reactions, combining a nucleophilic and an electrophilic half-reaction, with $\sum\Delta x = 0$.

The Grignard reaction on a ketone is shown in Figure 1 as two half-reactions with a net structural change only on the α-carbon of each half; the Michael reaction involves an addition, i.e., RH•RΠ.HΠ, changing at three adjacent carbons overall in two half-reactions, a nucleophilic RH, and an electrophilic addition RΠ.HΠ. The aldol reaction, or C-alkylation or acylation α to a ketone, is similarly just RH•RZ. The allylic alkylation of an enol ether (or enamine) is also shown and has the same designation as the Claisen rearrangement. Again, little more than 1% of the construction and fragmentation half-reactions examined had strands of more than three coupled atoms ($\alpha\beta\gamma$) out from the two skeletally changing α-carbons.

If all the unit reactions and half-reactions of Tables II and III are combined in one complete periodic table of reaction change it requires three dimensions for the three variables Δx, $\Delta\pi$, $\Delta\sigma$, with the construction half-reactions of Table III placed above ($\sum\Delta\sigma = +1$) the refunctionalizations of Table II ($\sum\Delta\sigma = 0$), and the fragmentations ($\sum\Delta\sigma = -1$) below those.

As with the refunctionalizations, actual constructions and fragmentations may consist of more than one unit reaction, some constructions being double ones like the Diels–Alder reaction, forming two C–C bonds, but much more commonly a combination of a single construction and a refunctionalization, known as a *composite construction*. Common examples include prior reduction to a carbanion (HZ) followed by construction RH, as in the Grignard half-reaction, or elimination following construction, as in the Wittig reaction (RH.RZ + ΠZ.ΠZ) or the half-reaction of carbonyl addition at α (RZ) followed by elimination at α,β (ΠZ.ΠH).

So far the only skeletal atoms of reference have been carbon with its valence of 4. However, the skeleton of a molecule may be seen as incorporating heteroatoms, usually N, O, S. These atoms in the skeleton may be described with exactly the same set of attachments and definitions as long as they are themselves understood to have a valence of 4, and this is simply arranged by assigning their unshared electron pairs as H, implying conjugate acid hy-

Table IV. Summary of the Reaction Databases

reaction type	ORGSYN	THEIL	CLF	SYNLIB	total
refunctionalization[a]	3660	34658	19810	32110	90238
classified	3334	31015	17552	29143	81244
unclassified	326	3643	2058	2967	8994
single construction	894	10080	11705	11239	33918
Classified (C–C)	605	5760	8922	7916	23203
classified (C–N)[a]	230	3522	1647	1798	7197
unclassified (C–C)[a]	59	798	1136	1525	3518
single fragmentation	52	923	703	1220	2898
classified	52	923	703	1220	2898
classified	16	576	420	777	1789
unclassified	36	352	283	443	1114
double construction	73	567	2103	1670	4413
multiple construction	52	390	600	246	1288
single rearrangement	33	594	572	1019	2218
complex combinations	78	946	1450	4722	7196
unmapped reactions	502	3593	4214	12473	20782
total reactions	5114	48229	39510	62901	155754

[a] Unclassified C–N constructions remain as refunctionalizations; classified C–N constructions also appear in refunctionalizations but are not counted twice in the total reactions columns.

drogens on the pairs. In our treatment of the databases below we accepted only all-carbon skeletons for refunctionalizations and so C–N bond formation is indexed as a refunctionalization; however, the formation of any C–N bond was separately treated also as a construction.[6] In these cases, with nitrogen as the α-atom in a C–N construction, the usual change at nitrogen (as nucleophile) would be a simple RH half-reaction, and so would usually have an electrophilic half-reaction (e.g., RZ) at the joining carbon, while with nitrogen as the β-atom, addition to an imine would be the electrophilic addition, RII.HII, the second (HII) change being that on the β-nitrogen.

Indexing Reaction Databases. In principle, any collection of reactions can be indexed by describing Δh, Δz, $\Delta\pi$, $\Delta\sigma$ at each changing carbon (or other skeletal atom).[7] This can be applied equally to books and catalogues of reactions as well as databases. The reactions can then be separated by $\sum\Delta\sigma$ for easier manipulation: into single constructions ($\sum\Delta\sigma = +2$ for the two half-reactions); refunctionalizations ($\sum\Delta\sigma = 0$); and single fragmentations ($\sum\Delta\sigma = -2$). The others are few, double constructions having $\sum\Delta\sigma = +4$, multiple constructions with more, etc. Skeletal rearrangements also have $\sum\Delta\sigma = 0$ but the absolute value, $|\sum\Delta\sigma|$, equals 2, and this is also tested to separate the rearrangements. Since $\sum\Delta\sigma$ is known in each group and Δh is available by difference we only require Δz, $\Delta\pi$ at each carbon to describe any reaction.

We applied this idea to the REACCS and SYNLIB databases.[8] These currently contain a total of about 155 000 reactions. The two databases apparently have little overlap.[9] In order to access their information it is necessary to have the mapping of the atoms from reactants to products. Given the atom mapping we can identify in each reaction entry the bonds made and broken. Since some entries are incompletely mapped, as discussed in the Appendix, the total number of accessible reactions is about 135 000.

This first breakdown of accessible entries in the databases affords some useful perspective on their contents (see Table IV). The largest group is the refunctionalizations, about 90 000, or two-thirds of the total reactions, and over 80 000 of these have only one or two unit reactions. The next largest group is single C–C constructions, something over 23 000; when the C–N heteroconstructions are added there are over 30 000 single constructions, almost a quarter of the total. There are less than 3000 fragmentations, and the total of other kinds of reactions show only about 4400 double constructions and some 10 000 others.

Refunctionalizations constitute the main group of reactions in the databases, with the unit reactions of Table II. The characteristic changes $\sum\Delta h$, $\sum\Delta z$, and $\sum\Delta\pi$ allow them all to be indexed and identified as single unit reactions at one site, two unit reactions at the same site, two unit reactions at two sites or more than two changes at one or more sites. Fortunately 90% of the mapped refunctionalizations are simple, having only one or two unit reactions.

These refunctionalizations can easily be written out for each of the possible unit reactions in Table II, cf., Figure 1. Thus, an allylic reduction [R′] would show a reacting strand of three adjacent carbons with $\Delta h = +1$ and $\Delta\pi = -1$ at the first carbon (labeled α in Figure 1, HII), $\Delta\pi = 0$ at the second (β) carbon (IIII), $\Delta\pi = +1$ and $\Delta z = -1$ at the third (IIZ). All possible combinations of two successive unit refunctionalizations anywhere on the same skeleton can then be written out as well (146 combinations from the 14 unit reactions of Table II). This results in 160 separate "storage boxes" for the refunctionalizations, which average about 500 reactions per box. The problem of pruning down the large number of hits by closer matching of structure at the reacting site is addressed in the next section. Some of the classified refunctionalizations will also be classified as C–N constructions, accepting the nitrogen as skeletal. These can then be retrieved from either classification. Those relatively few refunctionalizations with changes which do not match the 160 possible ones for one or two unit reactions are then discarded as too complex to classify.

The group of single C–C constructions is sorted according to the detailed functional changes in their half-reaction strands (Table III). These are classified following the procedure developed for our interface with the SYNGEN program.[10] Each entry is indexed with the characteristic change in z and π on each of the three atoms ($\alpha\beta\gamma$) on the two strands of the construction. The change index identifies each half-reaction as nucleophile or electrophile. We had previously defined and named 25 half-reactions for SYNGEN use, 16 nucleophiles and 9 electrophiles, including all the unit half-reactions of Table III, some common composites of those with unit refunctionalizations, and some further subdivision based on oxidation state level of the reactant. These categories were used here to identify the several construction half-reactions.

SYNGEN only accepts isohypsic[2] full constructions, i.e., a nucleophilic half-reaction combined with an electrophilic one, and these are by far the most precedented kind in the databases also. However, in the databases we also found some constructions to be nucleophile–nucleophile pairs, i.e., oxidative couplings, not accepted in the SYNGEN interface,[10] but now incorporated for retrieval here. We also discovered in our indexing a small number of new but

(6) Mapping of the O and S atoms was withdrawn from the current REACCS databases and so cannot be examined as skeletal atoms, although our original procedure was set up to include them.

(7) Indeed only three variables are needed but it is usually more convenient to record all four instead of having to determine one by difference from 4.

(8) The databases examined were: REACCS, version 8.0, consisting of ORGSYN (4735 entries), THEILHEIMER (46 784 entries), and CLF (Current Literature File, 32 742 entries), from Molecular Design Ltd., San Leandro, CA, and SYNLIB (62 901 entries) from Distributed Chemical Graphics, Inc., Meadowbrook, PA. Owing to multiple reaction products in some entries the total number of reactions is somewhat larger than the number of entries.

(9) Borkent, J. H.; Oukes, F.; Noordik, J. H. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 148–150.

(10) Hendrickson, J. B., Miller, T. M. *J. Am. Chem. Soc.* **1991**, *113*, 902.

moderately common composite half-reactions, and these have been added to the classifications. In a second iteration through the databases, taking both C and N as skeletal, the same procedure is followed to reclassify C–N bond formation from refunctionalization to construction and to index them the same way. The totals are seen in Table IV, in which the C–N formations are counted both ways but not duplicated in the total.

The construction entries are then stored in a matrix of nucleophile × electrophile half-reactions, so that each matrix element contains only the full constructions of one family and may be rapidly retrieved.[10] The classified entries constituted 90% of the constructions in the databases, about 65% being single unit reactions and 25% composites of two unit reactions. The unclassified constructions tend to be more complex constructions with several successive unit reactions, often involving subsequent aromatizations.

At this point it may be noted that many of these multistep reactions, with a number of unit reactions in succession, are entered in REACCS several times, spelling out the successive changes in three or four entries with the same reference. In the REACCS ORGSYN database,[8] for example, about a quarter of the entries were found to be such repeated references. The simple construction step entry may thus be separately identified and classified, and the unclassifiable entry with the complex overall transformation may be regarded as extraneous and ignored without loss.

The relative frequencies of the different construction half-reactions are of interest to organic chemists since they imply a measure of the scope and reliability of any construction reaction that might be chosen for a synthesis. The most frequently found nucleophiles are the enolate anion at 17% without subsequent elimination and another 8% with elimination, together about a quarter of all nucleophiles. A close second, nearly another quarter (23%), is the reductive carbanion typical of Grignard reagents. The next most common (17%) is the π-nucleophile, usually as the aromatic ring in aromatic substitution. Following these are the heteroatom-stabilized carbanions with 14%, including the Wittig nucleophile, a composite which forms a double bond by subsequent elimination after construction, and the enol ether/enamine family with 12%. The other nucleophiles are less than 10%.

Among electrophiles the ketone/aldehyde is the most common by far, either as simple carbonyl addition or as the composite addition–elimination (cf., Wittig reaction, etc.), a total of 42%. Following this are the simple alkylations with 16% and acylations with 14%. The conjugate additions are next with 11%, leaving only one-sixth of the total for the other electrophiles.

Single fragmentations are much less common and are handled just like constructions, since they may be seen as the reverse transformation (Table III). That there are so few compared with constructions reflects the much lower interest in these reactions among synthetic chemists.[11] The single rearrangements are recognized as one fragmentation and one construction with a carbon in common, and these too are few in number.[11] The double and multiple constructions are also relatively few and currently not classified.[12]

The overall survey of the databases (Table IV) shows that we can presently access some 30 000 constructions (including C–N), 1800 fragmentations and 74 000 refunctionalizations, a total of some 106 000 reactions.

The RETRIEVE Program. An index file was created for each of the entries in the several databases, using the reaction change parameters described above to classify each one. Then a program (RETRIEVE) was written to provide retrieval of all entries whose classified reaction change matched that of a given query. This constitutes a reliable procedure for locating all literature precedents in these databases for any reaction of interest. Its operation begins with the input on a graphics screen of a query reaction, drawn by the user. The facile and rapid drawing mode developed for the SYNGEN program has been adapted so that the user first draws the product(s) of a reaction as for a target molecule in SYNGEN, then on command the screen displays it twice with an arrow between and the user alters the left duplicate to create the reactant(s). The screen then displays a numbering of skeletal atoms on both sides to check the correspondence (mapping) of atoms; if it is incorrect, it may be changed.

The query so presented is first examined by the program and identified as single construction, fragmentation or refunctionalization, then classified into the family subset characterized by its changes in z and π at the reacting carbons. The screen then displays the query and its classification and requests a choice of database[8] to search, whereupon it displays the number of matching entries found in that database.

At this point there will usually be too many entries to examine and so the screen offers a set of options for the user to define a closer matching of the query structure with those of the database entries retrieved. With each matching option selected the screen displays the new, usually lower, total of matching entries until the number found is judged to be manageable for viewing. This pruning procedure is based on matching the values of the *unchanging* attachments on the reacting skeletal atoms, i.e., their values of σ, z, and π. For constructions these are offered in the succession α, β, γ out from the two joining atoms. The *nature* of the skeletal atom (C, N) is automatically matched for the α-atom but left for the user to match for the β- and γ-atoms. There is also offered a matching of inter- or intramolecular construction and whether the two joining (α) atoms are themselves already in rings in the reactants. For refunctionalization reactions, in which α, β, γ atoms are not explicitly defined, the query reaction is redisplayed with its reacting center atoms marked as α, β, γ to focus the pruning selections on each reacting atom separately.

The constructions sort themselves into the Nu × E matrixes evenly enough in number that the matching procedure is generally quite adequate to prune down quickly to a manageable number of hits, but with some refunctionalizations this is not the case. A frequency poll of the classified refunctionalizations[13] shows that the single unit reactions are the most common cases by far. Among these the simple isohypsic substitutions (ZZ, HH) are so abundant that they require another level of matching, i.e., by the atom type of the entering and leaving atoms in the substitutions, and this matching option is also offered to the user in these cases. Even so the few most common atom types (N and O) still predominate and require a further pruning basis.

We took from SYNGEN the categories used to define subsets of attached heteroatoms, z, generalized in terms of their *synthetic function*. The subsets are: L for leaving group, E for electron-withdrawing group, O for electron-

(11) Hendrickson, J. B. *J. Am. Chem. Soc.* **1986**, *108*, 678.

(12) A format for classifying the double constructions and rearrangements is currently under study and should in a future version be able to locate most of these.

(13) Hendrickson, J. B.; Miller, T. M. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 403.

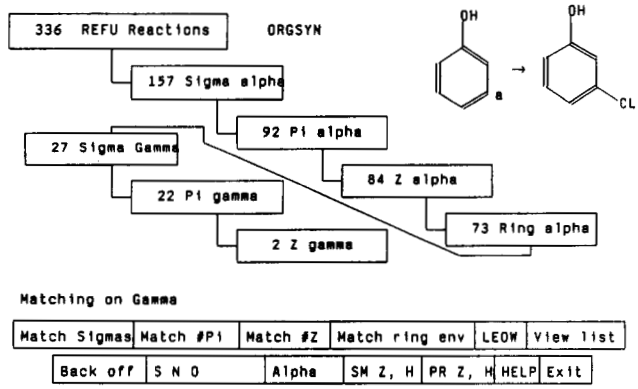

Figure 2. Pruning a refunctionalization reaction.

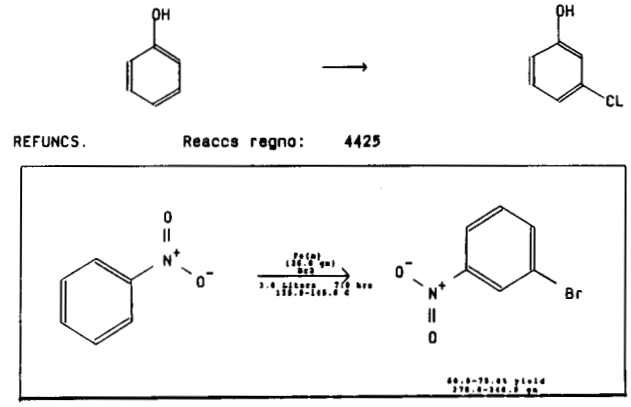

Figure 3. Matching entry for Figure 1.

donating group (for $x = 1$) and, for $z = 2$ or 3, W for carbonyl-type withdrawing function, to distinguish them from acetals, dihalides, etc. Thus matching by the "LEOW" nature of attached heteroatoms constitutes a further pruning option for any query. When this pruning has reached an acceptable number of hits, the user may command a sequential viewing of the matching entries found. When finished viewing, he returns to the database selection screen where he may elect another database to search.

To illustrate the procedure, a query is entered (shown at the top of Figure 2 with an a marking the reacting center) and is designated by the program as an oxidative refunctionalization ([X] in Table II). After establishing the atom correspondence, the screen displays Figure 2, showing the number of hits as 336 and a set of choices for pruning. As each choice is made from the menu below a new number of hits is displayed as shown, until the user is satisfied that a reasonable number (only 2 in this instance) has been reached; he then views the entries, one of which is shown in Figure 3.[14] Neither has the exact z-heteroatom requested, and so that particular case is recognized as not present in the database. The entry at the top of Figure 4 showed initially 3496 entries, a usual initial finding for such a common substitution. However, matching 12 choices of σ, π, z, and "ring" for the three atoms (α, β, γ) out from that bearing the chlorine rapidly reduces the choices to only seven entries, including the close match shown in Figure 4. In a construction case (see top of Figure 6) the pruning from 131 entries in the CLF database[8] locates finally only three cases (Figure 5), and again one is a close match (Figure 6). The two half-reactions are designated as R1.13 from the SYNGEN nomen-

(14) The actual REACCS entry also shows the literature reference below but this has been deleted in the figure to save space in reproduction.

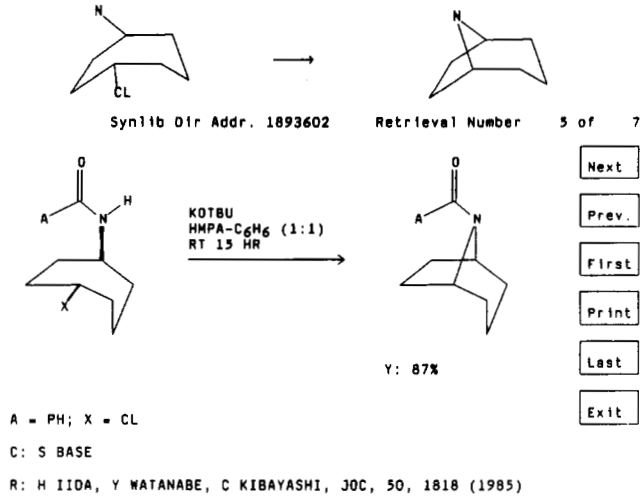

Figure 4. Matching entry for a substitution reaction.

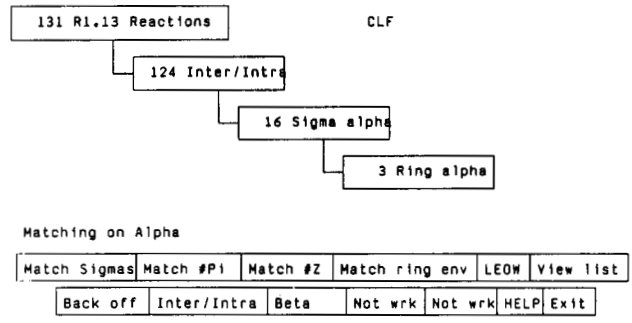

Figure 5. Pruning a construction reaction.

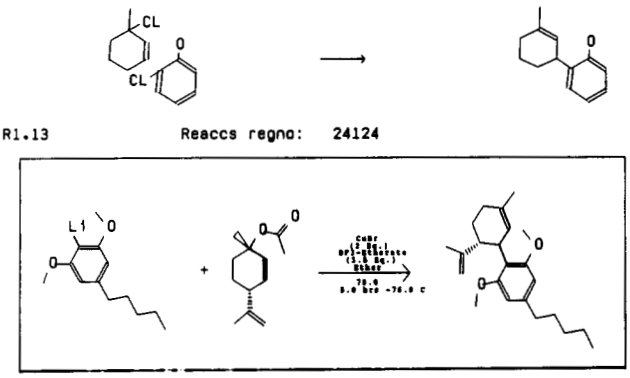

Figure 6. Matching entry for Figure 5.

clature:[10] R1 refers to a composite reductive nucleophile, HZ + RH, and 13 refers to the allylic electrophile, RII. IIII.IIZ on $\alpha\beta\gamma$.

In summary, the system of reaction description is simple and fundamental. It meets the initial requirement that any reaction have a defined place in the system. Furthermore, the system also provides for the easy pruning of large groups of reactions of a general class to small groups of more closely refined similarity. The success of the RETRIEVE program in finding correct matches from the databases not only lends confidence in organizing reactions in this way, but also the program itself is fast and facile for practical searches. The literature precedents obtained for a query are sharply defined so that it is always clear just what one will get, and what will be excluded.

The RETRIEVE program is available (coupled with SYNGEN under a single executive) on application to the authors. It operates on a VAX or microVAX computer under VMS, with Tektronix-style graphics and is automatically linked to the REACCS and/or SYNLIB databases.

Appendix. Computing Details. The first problem in accessing the databases is the atom mapping between reactant and product structures.[13] Some 20 000 entries have incomplete mapping of the carbon atoms and so are usually inaccessible to our application. If carbons in the product have no correspondence to those in the reactant, i.e., are unmapped, they represent another reactant annotated only as text over the reaction arrow, cf., formaldehyde, methyl iodide, acetylene, etc. If any product carbon is engaged in a construction and is unmapped because of such a text source, the reaction is discarded.[15] If any unmapped carbon is only involved in a refunctionalization, however, it may be classified. Thus, if the carbons of an acetate replacing a chlorine are not mapped, the change is simply indexed as C–Cl → C–OR. Conversely, when a carbon in the reactant does not appear in the product it will be accepted for a refunctionalization (or construction) of the mapped atoms (as in loss of an unmapped protecting group). It will, however, cause rejection of a fragmentation if one of the unmapped carbons is cleaved from a mapped one. The available mapping in SYNLIB[8] was often inadequate; this was overcome by converting the SYNLIB entries into the REACCS RD-file format first in order to take advantage of the REACCS automatic atom mapping routine.[13]

The following procedure is used to classify each database entry.[13] At the start only carbon is accepted as skeletal. Entries with unmapped product carbons are only examined if no mapped C–C bond is altered, i.e., refunctionalization. The other entries are divided into reaction categories according to the C–C bonds made and broken, as described above.[16] If nitrogen is present the program then proceeds to a second iteration in which nitrogen is accepted as a skeletal atom along with carbon. There follows a parallel sorting, and deletion of entries with the same classification as found in the first pass. In the refunctionalizations category *changes at skeletal nitrogen* (cf., oxidation–reduction), which would have been classified as no change in the carbon skeleton iteration, are now labeled.

In documenting each entry the values of Δh, Δz, $\Delta\pi$, $\Delta\sigma$ are recorded for each changing skeletal atom,[10] as well as their values of h, z, π, σ in the reactant, for use in pruning. Also recorded for pruning is the presence in a ring of a reactant atom and if a new ring is formed in a construction. For constructions and fragmentations the two changing $\alpha\beta\gamma$-strands out from the forming/breaking skeletal bond are each described by a $\Delta z\pi$-list, defined as $(z_\alpha\pi_\alpha z_\beta\pi_\beta z_\gamma\pi_\gamma)_{REAC} - (z_\alpha\pi_\alpha z_\beta\pi_\beta z_\gamma\pi_\gamma)_{PROD}$. Since only two bits are needed in the computer to record either z(0–3) or π(0–2), this $\Delta z\pi$ list only requires 12 bits to describe a half-reaction.[10] This list serves as a simple identifying number for each half-reaction type. The two $\Delta z\pi$ lists for the two half-reactions now fully classify any single construction or fragmentation. In most full construction reactions found to be unclassifiable (Table IV), only one of the half-reactions is found to be unclassifiable while the other is normal and classifiable.

For refunctionalizations, in which no $\alpha\beta\gamma$-strand is defined, a different key for the changes is used. This includes the *number* of skeletal atoms changing as well as Δz and $\Delta\pi$ lists, each list consisting of the numbers of atoms which change by −3, −2, −1, +1, +2, +3 in z or π. As noted above there are 160 refunctionalizations of 1–2 unit reactions and their keys were first all generated by hand from the net structural change in the chemistry of each one.

Taken together these reaction keys constitute unique ID numbers for each kind of reaction, characterizing it differently from any other kind. The number of entries for each ID number is kept for fast display in the retrieval searches (see Figures 2 and 5). In addition the entries themselves are grouped by their ID numbers for fast pruning and display. Finally, when a query reaction is entered by the user it is classified in the same way, to provide an ID number which may then be matched to those in the databases.

The RETRIEVE program also contains the modules which are used to organize any other reaction database by first converting its entries into the REACCS RD-file format for mapping and then indexing the entries according to the groupings of Tables II and III as discussed above. This not only allows the program to be used to index and search other databases, e.g., proprietary ones, but also allows statistical analyses of the breakdown of the reactions contained within them, as in Table IV, or in more detail by reaction type.

Acknowledgment. We are grateful to the National Science Foundation (Grant CHE-8620066) and to the Eastman Kodak Company for generous support of this work.

(15) In many cases, especially with C–N heteroconstructions, the reactant will have been a simple one and can be deduced from the nature of the product. Expansion to incorporate cases not now accessible is planned for future development.

(16) In about 300 cases more than one β or γ atom is changed in one half-reaction. Entries with these multiple changing strand atoms were discarded as too complex; they generally involve concurrent changes unrelated to the construction itself, as with an attendant aromatization.