COGNOS: A Beilstein-Type System for Organizing Organic Reactions

James B. Hendrickson* and Thomas Sander

Department of Chemistry, Brandeis University, Waltham, Massachusetts 02254

Received December 21, 1994[⊗]

We describe a logical system to organize and index organic reactions. This is implemented in the COGNOS program to search reaction databases for literature precedents. It is based on the *net structural change* in a reaction, rather than on substructure searching. This system rigorously provides a place for any possible reaction, generalizing and classifying all reactions into families by their changes in skeleton and functionality. These families are digitally identified, allowing databases to be indexed and grouped in numerical order so as to afford instant retrieval of matching reactions. The retrieved entries are further pruned by structural details at the reacting atoms to provide small, manageable sets most closely matching any input query.

Since its inception organic chemistry has never developed a comprehensive, rigorous system for classifying and annotating reactions. Now that reaction databases have appeared for computerized retrieval, the need for such a system of organization has become acute. The organization of reactions should mirror the Beilstein system for organizing compounds, in having a limited number of main categories and a defined hierarchic, or taxonomic, nesting of subfamilies within them to further refine lesser distinctions. The main criterion is that the system must be capable of describing all possible reactions, whether currently known or unknown, so that any possible reaction has a clear place in the organizational scheme. Such a system must be capable of defining precisely but in generalized terms the *net structural change* in any reaction, i.e., the bonds made and broken from substrate(s) to product(s).

We outline first a comprehensive system that meets these criteria for the organization of reactions. Following this is a description of the COGNOS program which implements it for the retrieval of reactions from large databases. Very rapid retrieval is made possible by indexing and arranging all database entries according to this organizational scheme, so that all entries of the same kind are located together.

Central to this development is the importance of carbon in organic chemistry. All structures are viewed here as comprising a framework or skeleton of carbon atoms linked by σ-bonds, with functional groups—as attached heteroatoms or C−C π-bonds—located at specific sites on these carbon skeletons. This dichotomy of skeleton and functionality for the *structure* of compounds is paralleled in their *reactions* by *constructions* and *fragmentations,* reactions which construct or cleave the C−C σ−bonds of the skeleton, and *refunctionalizations*, reactions which alter functional groups without changing the skeleton.

A suitable definition generalizing and simplifying the description of the bonding to skeletal carbon has been available for some time,[1−3] and is summarized here. In this definition there are four synthetically important kinds of attachment or bonding which any single carbon may have

R for σ-bond to another carbon (skeletal bond)

Π for π-bond to another carbon (functional bond)

Z for bond (σ- or π-) to electronegative heteroatom (N, O, S, halogen, etc.)

H for bond to hydrogen or electropositive atom (B, Al, Si, Sn, metal, etc.).

The number of bonds of each kind is then defined as σ, π, z, h, respectively, with a sum of 4. The heteroatom attachments are distinguished as electronegative and electropositive to afford recognition of the oxidation state at that carbon by $x = z - h$, with values of $-4 \le x \le +4$. The oxidation state change in a reaction is the sum of Δx over all changing carbons.

Definition of Reactions. The essence of a chemical reaction is the exchange of attachments (or bonds) on each changing carbon in the reaction. The simplest reaction change is a single exchange of one attachment for another on one carbon, i.e., one bond made and one bond broken. On any one carbon there are 16 possible such single exchanges which derive from these four attachment types. We may label these changes at any carbon with a simple notation of two letters, the first for the bond made, the second for the bond broken, as shown in Table 1; the changes in the attachment types are shown as Δx, $\Delta\pi$, and $\Delta\sigma$.

These simple exchanges are readily recognized as the familiar reaction families of *substitution, elimination/addition*, and *construction/fragmentation.* Each has an oxidative and reductive variant: any carbon changed by +H or −Z is reduced, and any by −H or +Z is oxidized, each by one oxidation state level. Thus the simple redox substitutions themselves, HZ and ZH (i.e., +H −Z and +Z −H, respectively), change by two levels of oxidation state. The former, HZ, is a reductive substitution of a heteroatom bond by H, as in ketones or halides by hydride; the latter, ZH, is an oxidative substitution of H by heteroatom Z, as in alcohol oxidation or aromatic nitration. The other substitutions are those of electrophiles (proton or metal exchange, HH) and of nucleophiles, as in S_N2 displacements by nucleophiles (ZZ).

A single exchange ΠΠ at carbon is a double bond shift, as in the central carbon of an allylic substitution, eq 1. The exchange RR at one carbon is characteristic of the migrating carbon in a simple 1,2-rearrangement, which makes one and breaks another C−C σ-bond as in eq 2; the RR exchange is not restricted to these rearrangements, but others are rare.

In Table 1 the four single exchanges with just H and Z can be seen to involve only one carbon overall, but in the rest changes in R or Π must involve the same change on an

[⊗] Abstract published in *Advance ACS Abstracts,* March 1, 1995.

0095-2338/95/1635-0251$09.00/0

Scheme 1

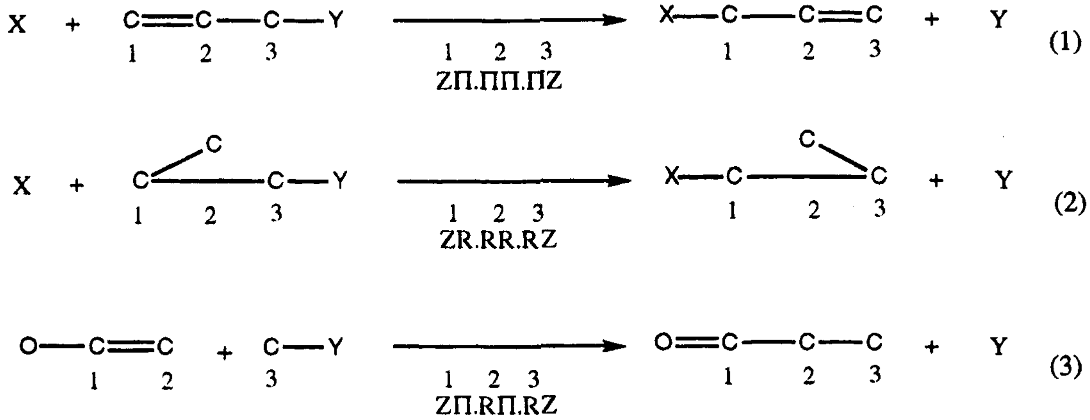

X + C=C—C—Y → X—C—C=C + Y (1)
1 2 3
ZΠ.ΠΠ.ΠZ
X + C—C—Y → X—C—C + Y (2)
1 2 3
ZR.RR.RZ
O—C=C + C—Y → O=C—C—C + Y (3)
1 2 3
ZΠ.RΠ.RZ

Table 1. The 16 Possible Unit Exchanges at any Skeletal Carbon

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

adjacent carbon, linked to it either in the substrate or product, or both, as in eqs 1 and 2. Such carbons have nonzero values of $\Delta\sigma$ or $\Delta\pi$ in Table 1. Thus, for $\Delta\pi$, a simple isohypsic[4] elimination is ΠH•ΠZ ($\Sigma\Delta x = 0$) on two adjacent carbons, forming a π-bond between them at the expense of H on one and Z on the other, cf., dehydrohalogenation. An oxidative addition to a double bond, as with bromine, is ZΠ•ZΠ ($\Sigma\Delta x = +2$ from Table 1) and reductive hydrogenation of a double bond would be HΠ•HΠ with $\Sigma\Delta x = -2$. Exchanges at three carbons are characteristic of allylic, or vinylogous, substitution reactions, the middle carbon changing ΠΠ, as in eq 1 or the allylic reduction HΠ•ΠΠ•ΠZ, which indicates the reaction H + C=C−C−X → H−C−C=C + X. In these reactions the first two carbons are linked by −Π, the last two by +Π. All of these reactions are refunctionalizations; no skeletal C−C σ-bond is altered.

Reactions which alter the skeleton must also involve two carbons, linked in the product (constructions, +R) or in the substrate (fragmentations, −R). A simple construction, as organolithium reaction with a ketone, is RH•RZ, linking two adjacent carbons with +R. This is read as a construction reaction, which forms a C−C σ-bond (R), with loss of H (i.e., Li) at one carbon and loss of Z (i.e., bond to oxygen) at the other carbon. There are more than two carbons involved in constructions like the Michael addition (C−H + C=C → C−C−CH), described by RH•RΠ•HΠ, in which the left two carbons are linked by +R, the right two by −Π. A typical rearrangement would be ZR•RR•RZ as in eq 2, with three adjacent carbons linked either in substrate or product, fragmenting one skeletal bond and constructing another; the left two carbons are linked with −R and the right two with +R. A notation parallel to eqs 1 and 2 can be made for the three-carbon change in enolate alkylation, shown as eq 3; the middle carbon is now RΠ so that the 1−2 bond is an addition (ZΠ•RΠ) and the 2−3 bond is a construction (RΠ•RZ).

Unit Reactions. We may now define a *unit reaction* as a single exchange of attachments at each changing carbon, and it will be apparent that all carbons in a unit reaction are adjacent, linked carbons, annotated either as a linear string like those above, or a cyclic one, for which the notation may be distinguished with parentheses. Photolytic cycloaddition will then be (RΠ•RΠ•RΠ•RΠ) on the four carbons in a four-ring cycle, and the Diels−Alder cycloaddition is (RΠ•RΠ•ΠΠ•ΠΠ•RΠ•RΠ) as a six-ring cycle. (The general representation of reactions as cycles is treated in ref 3.) In the linear strings the linked pairs of carbons in any unit reaction are all connected by sharing either ±R or ±Π. Only the end carbons of the string will have ±H or ±Z.

Skeletal alteration reactions are treated as independent *half-reactions.*[1] The simple construction or fragmentation unit reactions are divided into two unit half-reactions,[1−3] showing the reaction changes on each of the two strands of carbons out from the two carbons which form (or break) the skeletal C−C σ-bond. These represent the changes on each of the two substrates in an intermolecular construction or the two products of a fragmentation (In intramolecular constructions/fragmentations the half-reaction notation is the same even though the two strands are actually linked at their outer ends.). The notations for half-reaction strands always start with ±R and end with ±H or Z. The organolithium + ketone construction above has the two half-reactions RH and RZ. The right-hand, electrophilic, half-reaction in the Michael addition above is just RΠ•HΠ, while the left half-reaction, RH, is the nucleophilic half of the construction.

The more complex double constructions and fragmentations and the rearrangements are unit reactions characterized by *two* ±R carbons. Double construction half-reactions on two carbons will be RΠ•RΠ or the vinylogous RΠ•ΠΠ•ΠΠ•RΠ on four carbons; the reverse fragmentations are ΠR•ΠR or ΠR•ΠΠ•ΠΠ•ΠR. The two double construction half-reactions taken together will characterize a full unit reaction: the double construction of the Diels−Alder reaction above. The two fragmentations together describe its reverse, the retro reaction.

$$\mathrm{R{-}Cu + C{=}C{-}COOEt + R'{-}Br \rightarrow R{-}C{-}C(R'){-}COOEt} \quad (4)$$

In noncyclic cases the overall unit reaction for these double changes will consist of three half-reactions: two single construction or fragmentation halves at each end of a double one. Thus the cuprate addition to an unsaturated ester followed by alkylation of the enolate shown in chemical shorthand as eq 4, will be overall RH•RΠ•RΠ•RZ, seen as three half-reactions, i.e., RH + RΠ•RΠ + RZ. Rearrangements will also have three half-reactions, the central one being RR as in eq 2, or possibly the vinylog RΠ•ΠΠ•ΠR, flanked by a single fragmentation at one end and a single construction at the other.

The central dichotomy of skeleton and functionality is now invoked to classify reactions first by their changes in skeleton and then into families of functionality change within each skeletal class. These seven basic classes of skeletal change can be established first:

Refunctionalization Class:	Refunctionalizations at carbons
	Refunctionalizations only at heteroatoms (below)
Skeletal Alteration Class:	Single Constructions
(half-reactions)	Single Fragmentations
	Double Constructions
	Double Fragmentations
	Rearrangements

An analysis of four major commercial reaction databases (155 000 reactions) described the proportions in each class;[5] almost three-quarters of the entries are refunctionalizations, and most of the rest are single constructions. This analysis also showed that about 80% of all entries are in fact simple unit reactions, and about two-thirds of the those remaining are just two successive unit reactions, such as the Wolff–Kishner reduction, described as HZ + HZ on one carbon, or the Wittig reaction as RH•RZ + ΠZ•ΠZ on two carbons. These reactions involving two successive unit reactions on the same carbons are called *composite* reactions. It was also clear in the survey[5] that the number of linked carbons is rarely more than four and almost never more than six, as in the Diels–Alder reaction.

Generation of Reaction Families. We can now logically generate all the possible unit reactions taking place over a string of adjacent changing carbons. This may be done by stringing together all valid combinations of single carbon exchanges from Table 1: any carbon with +Π or −Π must have an adjacent carbon with the same change, as must any with +R or −R. There are two major categories of reactions distinguished above: a refunctionalization class with no ±R and the skeletal alteration classes (with ±R). Each class contains a set of reaction families which define the possible functionality changes in the class. These changes may be described in terms of the change per carbon in Table 1. We see at work in these combinations of changes the three familiar variables $\Delta\sigma$, $\Delta\pi$, and Δx, i.e., construction/fragmentation, addition/elimination, and oxidation/reduction.

(a) Refunctionalizations. The refunctionalization class, with $\Sigma\Delta\sigma = 0$, involves only two of these variables and so may be organized by $\Sigma\Delta\pi$ and $\Sigma\Delta x$, as in Table 2, which shows the 10 families of simple refunctionalizations. The refunctionalization families are simply characterized with familiar labels, shown in brackets for each family in Table 2. These are [X] for oxidative, [R] for reductive, [A] for addition, [E] for elimination, and [S] for substitution; H-substitution is separated as [H]. These unit reactions change either one carbon or two adjacent carbons.

Table 2. Unit Refunctionalization Reactions

$\Sigma\Delta x =$	−2	0	+2	
$\Sigma\Delta\pi =$				
+2	ΠZ·ΠZ **[RE]**	ΠH·ΠZ **[E]**	ΠH·ΠH **[XE]**	Elimination
0	HZ **[R]**	HH, ZZ **[H], [S]**	ZH **[X]**	Substitution
−2	HΠ·HT **[RA]**	HΠ·ZΠ [A]	ZΠ·ZΠ **[XA]**	Addition
	Reductive	Isohypsic	Oxidative	

Another 10 families may be created in parallel using their vinylogs, adding two more carbons into each strand as a π-bond. Thus the simple substitution (ZZ on one carbon) becomes allylic substitution (ZΠ•ΠΠ•ΠZ) on three carbons, as in eq 1, and the four-carbon vinylog of reductive addition will be HΠ•ΠΠ•ΠΠ•HΠ, a 1,4-addition of hydrogen. The 10 vinylogs are now designated with primes, as [S′] for the allylic substitution above (ZΠ•ΠΠ•ΠZ; eq 1), and [RE′] and [XA′], etc., for the four-carbon vinylogs of the eliminations/additions, i.e., [RA′] for the 1,4-hydrogenation HΠ•ΠΠ•ΠΠ•HΠ. There are also some doubly allylic reactions in the databases; these are all treated the same way, i.e., five-carbon doubly allylic substitutions [S″] and six-carbon doubly allylic eliminations/additions [E″], [A″], etc. Therefore, there are overall 30 families of unit refunctionalization reactions including vinylogs and double vinylogs, with reacting strands of up to six carbons.

(b) Skeletal Alterations. The skeletal alteration class involves the third variable, $\Sigma\Delta\sigma$, but since these reactions are treated as half-reactions, they may also be organized just by $\Sigma\Delta\pi$ and $\Sigma\Delta x$, assuming the $\Delta\sigma$ to be on the first carbon. These half-reaction families are shown in Table 3, which includes all possible unit half-reactions on one or two carbons. Each fragmentation is the reverse of a construction and has an inverted oxidation state change ($\Sigma\Delta x$) as well as the reverse $\Sigma\Delta\pi$. The eight construction/fragmentation half-reactions are similarly labeled (in brackets) by adding the letter C for construction or F for fragmentation to the reaction label. In the example of Michael addition above, one half-reaction is the reductive addition of an alkyl to an electrophilic double bond, i.e., R + C=C → R−C−CH, and would be labeled an [RAC] half-reaction, annotated in Table 3 as RΠ•HΠ with $\Sigma\Delta x = -1$; its nucleophilic other half will be RH (labeled as an [XC] half-reaction) with $\Sigma\Delta x = +1$. The full construction is therefore labeled as [RAC•XC] and is isohypsic overall ($\Sigma\Delta x = 0$).

The eight vinylogs of these simple half-reactions (on 3−4 carbons) are also labeled with primes and doubly vinylogous ones (5−6 carbons) with double primes as above, making 24 families of half-reactions overall on half-reaction strands of 1−6 carbons. The full constructions/fragmentations, made by combining two half-reactions, can therefore in principle have strands of 2−12 carbons, but those above five or six carbons are very rare.

The 54 families of Tables 2 and 3 with their vinylogs could all be merged and presented in a single table of all unit reactions with a third dimension for $\Sigma\Delta\sigma$, i.e., with single construction half-reactions ($\Sigma\Delta\sigma = +1$) as a plane above refunctionalizations ($\Sigma\Delta\sigma = 0$) and the fragmentation halves ($\Sigma\Delta\sigma = -1$) below. The double construction and fragmentation half-reactions ($\Sigma\Delta\sigma = +2$ and −2, respectively) would

Table 3. Single Construction/Fragmentation Half-Reactions

	Reductive		Oxidative		
$\Sigma\Delta x =$	−1	−1	+1	+1	
$\Sigma\Delta\pi =$					
+2	ΠR·ΠZ **[REF]**		ΠR·ΠH **[XEF]**		Elimination
0	HR **[RF]**	RZ **[RC]**	ZR [XF]	RH **[XC]**	Substitution
−2		RΠ·HΠ **[RAC]**		RΠ·ZΠ **[XAC]**	Addition
	Fragmentation	Construction	Fragmentation	Construction	

then be above and below the single ones on such a unified table, which represents a kind of "periodic table" of organic reactions.

All these families are *single unit reactions* and constitute most of any database.[5] The remaining reactions are usually composites of just two successive unit reactions. These can be generated rigorously by taking all combinations of two unit reactions with overlapping strands of carbons, as outlined in the next section. These would include, for example, reduction of acid to primary alcohol, two successive HZ exchanges on one carbon which would be labeled as [R + R] from Table 2. The two unit reactions in a composite construction (or fragmentation) are usually a construction composed of two half-reactions for one and a refunctionalization for the other, as with the Wittig reaction above.

In reactions which change *only* the heteroatoms in a functional group, bonds to carbon are unaffected. In these cases the heteroatoms themselves are treated as if they were carbon with four attachments:[6] charges are ignored and unshared electron pairs are annotated as H, i.e., as the conjugate base of an acid. Thus the common reactions are reduction [R] and oxidation [X], as with carbon, and they are labeled with the symbol of the changing base atom, as [R(N)] for reduction of nitro, or [X(S)] for oxidation of sulfide.

Classification Examples. Examples of common reactions and their reaction notations are shown in Figure 1. In the first one, three adjacent carbons are seen to change their attachments, and these three are numbered 1, 2, 3. Their changes are characterized below the arrow and are recognized as an allylic reduction, labeled as [R′]. The second case is similar, an allylic substitution [S′], but the difference from the first case emphasizes that the classification of reaction families is based strictly on the changes at the reacting carbons and not on the nature of the overall structure with its unchanging structural parts.

The next two cases in Figure 1 are Michael additions, constructions composed of two half-reactions, the electrophilic half the same in each case (RΠ·HΠ). In the first, four carbons change attachments, i.e., two in each construction half-reaction. The nucleophilic half (carbons 1,2) is an oxidative addition [XAC], the electrophilic half a reductive addition [RAC] at carbons 3,4. The chemically similar second case (4) has an *overall* net structural change of only RH, at carbon 1, in the nucleophilic half, and the electrophilic half is the same as in the previous case. The generalized attachments at the oxygen-bearing carbons are unchanged: the enol ether remains as $z,\pi = 1,1$ and the ester remains as $z,\pi = 3,0$.

The last two examples in Figure 1 show composite reaction notation. The Wittig reaction in case 5 has two joining half-reactions followed by a reductive elimination and so is labeled as [RC·XC + RE]. Another composite construction (case 6) is the joining of two simple half-reactions (carbons 2,3) followed by isohypsic[4] elimination at carbons 1,2 to aromatize the pyrone, hence labeled as [RC·XC + E].

Indexing the Databases. The key to rapid retrieval of reactions from large databases is to group them together by reaction families, providing an ordered index of all the entries. Then to arrange the reactions in numerical order in such an index, these families need digital identification. Each reaction family can be identified by returning to the definition of a reaction as the *net structural change* between substrate and product. The class of the reaction, i.e., the skeletal change, is determined first by recognition of any carbons with $\Delta\sigma \neq 0$. *Within each class* then the net structural change is just the change in the functional groups.

For any carbon in the skeleton of substrate or product the σ−value is known and, since $h = (4 - \sigma) - (z + \pi)$, the values of z and π at each carbon define its functionality. A $z\pi$-value for each carbon in a reacting strand can be defined with four bits: two bits for z (=0−3) and two bits for π (=0−2). Thus a $z\pi$-list is a list of these $z\pi$-values over the linked changing carbons of a reacting strand. If the $z\pi$-list of the product strand is subtracted from the corresponding $z\pi$-list of those carbons in the substrate, we obtain a $\Delta z\pi$-list. This is a binary number which serves as an identification number for the reaction family. At four bits per carbon, a strand of up to eight changing carbons requires just one 32-bit word for this reaction identification number.[7]

In order to classify a reaction the carbons in substrate and product are identified from their connection tables and assigned their σ, π, z, and h values. For the computer to do this, it is critical that the entry data have atom−atom mapping, identifying each carbon in the substrate with one in the product. The changing carbons are those which change any of their σ, z, and π values. The class of the reaction is first determined by the overall skeletal changes: refunctionalizations by $\Sigma|\Delta\sigma| = 0$; the others as the several skeletal alterations listed above depending on the number and nature of the several $\Delta\sigma$ carbons. In the latter classes the half-reaction strands are identified as changing carbons which remain σ-bonded throughout the reaction, and each half-reaction is defined separately. It may be noted also that in any reacting strand of more than one carbon all skeletal (C−C) bonds change their bond order in the reaction.

The reaction itself is then identified by its $\Delta z\pi$-list, obtained by subtracting the product from the substrate $z\pi$-lists.[7] With refunctionalization reactions the $\Delta z\pi$-lists can be read from one direction or the other along the reacting strand, i.e., from left to right or right to left. Of the two possible $\Delta z\pi$-list numbers the larger number is used as the

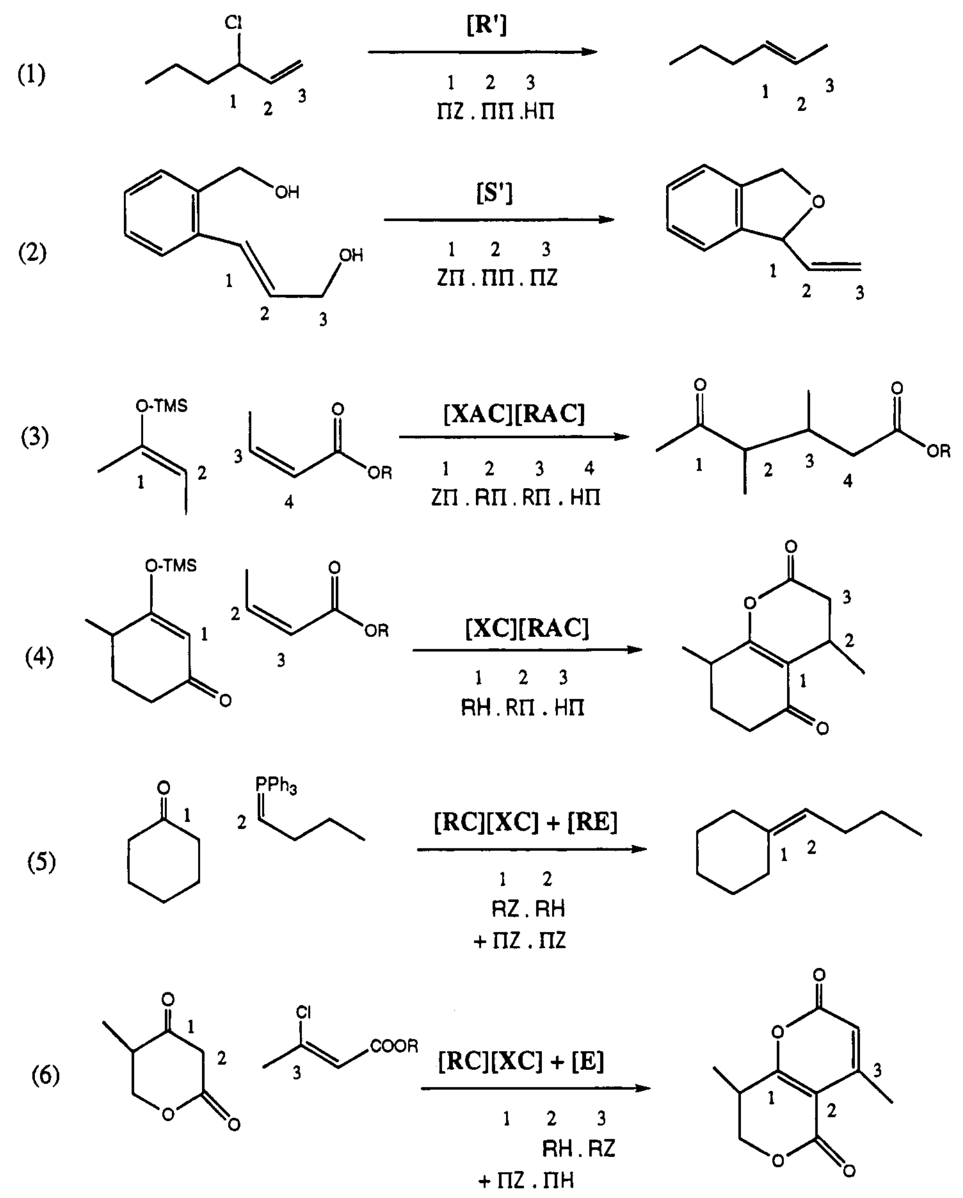


Figure 1. Examples of reaction classification.

identification number for the reaction. In skeletal alteration half-reactions the direction of the $\Delta z\pi$-list is uniquely defined as out the strand from the carbon which makes or breaks the skeletal bond, i.e., the one with $\Delta\sigma \neq 0$. In practice we found that there is essentially no duplication among these identification numbers for either the single unit reactions or the composites.[8]

To illustrate the process, examine the first reaction, [R′], in Figure 1, which is ΠZ•ΠΠ•HΠ for carbons 1, 2, 3. In the substrate carbon 1 has $z = 1$ and $\pi = 0$, hence a $z\pi$-value of 0100; carbons 2 and 3 have only $\pi = 1$, i.e., a $z\pi$-value of 0001. The whole process identifying the reaction is shown below and leads to the preferred identification number of **301** for the [R′] reaction. This is a hexadecimal number with one digit for each changing carbon.[7]

	Left to right:	Right to left:
C no.	1 2 3	3 2 1
	0100.0001.0001 (SUB)	0001.0001.0100 (SUB)
	−0001.0001.0000 −(PROD)	−0000.0001.0001 −(PROD)
	0011.0000.0001 $\Delta z\pi$-list (**301**)	0001.0000.0011 $\Delta z\pi$-list (**103**)

In the construction half-reaction [XAC] of case (3) in Figure 1, the change is RΠ•ZΠ at carbons 2,1, the construction carbon written first. The identification number arises from substrate − product = 0001.0101 − 0000.1000 = 0000.1101, or the hexadecimal number **0D** at the two carbons. The other half-reaction, [RAC], is similarly identified as 0001.0001 − 0000.0000 = **11**. All [XAC] half-reactions will be RΠ•ZΠ no matter how the changing carbons are substituted and will have a $\Delta z\pi$-list identifier of **0D**. This can be checked as above with a case at lower oxidation state, $Et_2C{=}CH_2 + CH_3I \rightarrow Et_2CZ{-}CH_2{-}CH_3$ and this will give the same identifier, **0D**, for the change on the alkene half-reaction. The methyl iodide half-reaction is RZ, a reductive construction [RC] in Table 3. The identifier for this is **4** as it is also for addition to a ketone or at the next higher oxidation state to a carboxylic acid derivative.

Every reaction or half-reaction has of course a reverse reaction in which all three variables are opposite: oxidation/reduction (Δx); addition/elimination ($\Delta\pi$); construction/fragmentation ($\Delta\sigma$). In Tables 2 and 3 the reverse reactions are centrosymmetrically oriented, as they would all be in a full three-dimensional table such as outlined above. Thus the reverse reaction of reductive addition [RA] in Table 2 is oxidative elimination [XE], and in Table 3 the reverse of an oxidative addition construction half-reaction [XAC], i.e.,

Table 4. Identifiers (ID) for Unit Reaction Families

		refunctionalization		construction		fragmentation	
strand n =	reaction type	$\Delta\sigma = 0$	ID	$\Delta\sigma = +1$	ID	$\Delta\sigma = -1$	ID
1	substitutions	[S], [H]	**0**				
	reductive	[R]	**4**	[RC]	**4**	[RF]	**0**
	oxidative	[X]	$\underline{\mathbf{C}}$	[XC]	**0**	[XF]	$\underline{\mathbf{C}}$
2	additions	[RA]	**11**	[RAC]	**11**		
		[A]	**0D**				
		[XA]	$\underline{\mathbf{CD}}$	[XAC]	**0D**		
	eliminations	[RE]	**33**			[REF]	$\underline{\mathbf{F3}}$
		[E]	**2F**				
		[XE]	$\underline{\mathbf{EF}}$			[XEF]	$\underline{\mathbf{EF}}$
3	allylic substitutions	[S′]	**2FD**				
		[H′]	**0FF**				
		[R′]	**301**	[RC′]	**103**	[RF′]	$\underline{\mathbf{F01}}$
		[X′]	$\underline{\mathbf{EFD}}$	[XC′]	**0FF**	[XF′]	$\underline{\mathbf{EFD}}$
4	vinylogous additions	[RA′]	**1001**	[RAC′]	**1001**		
		[A′]	**0FFD**				
		[XA′]	$\underline{\mathbf{CFFD}}$	[XAC′]	**0FFD**		
	vinylogous eliminations	[RE′]	**3003**			[REF′]	$\underline{\mathbf{F003}}$
		[E′]	**2FFF**				
		[XE′]	$\underline{\mathbf{EFFF}}$			[XEF′]	$\underline{\mathbf{EFFF}}$
5	double allylic substitutions	[S″]	**0FFFF**				
		[H″]	**0FFFF**				
		[R″]	**30001**	[RC″]	**10003**	[RF″]	$\underline{\mathbf{F0001}}$
		[X″]	$\underline{\mathbf{EFFFD}}$	[XC″]	**OFFFF**	[XF″]	$\underline{\mathbf{EFFFD}}$
6	double vinylogous additions	[RA″]	**100001**	[RAC″]	**100001**		
		[A″]	**0FFFFD**				
		[XA″]	$\underline{\mathbf{CFFFFD}}$	[XAC″]	**0FFFFD**		
	double vinylogous eliminations	[RE″]	**300003**			[REF″]	$\underline{\mathbf{F00003}}$
		[E″]	**2FFFFF**				
		[XE″]	$\underline{\mathbf{EFFFFF}}$			[XEF″]	$\underline{\mathbf{EFFFFF}}$

RΠ•ZΠ, is a reductive elimination fragmentation [REF], i.e., ΠR•ΠZ, in which all three variables are seen to be reversed.

For half-reactions, ordered out from the skeletal-change carbon, the identifier of the [XAC] half-reaction is **0D**, and the reverse reaction [REF] is $\underline{\mathbf{F3}}$, i.e., **−0D**.[7] For refunctionalizations, however, there are two possible directions for the strand, each with a $\Delta z\pi$-list, and only the larger is kept as the identifier. The two $\Delta z\pi$-lists for the reverse reaction will be the negatives of these, but since only the higher number is kept as reaction identifier in each case, the identifier of one may not be the negative of the other. The two $\Delta z\pi$-lists for addition [A] are **0D** and **−2F** ($\underline{\mathbf{D1}}$)[7] and for elimination [E] they are **2F** and **−0D** ($\underline{\mathbf{F3}}$)[7]. Thus **0D** is retained as the identifier for [A] and **2F** is kept for [E]. The hexadecimal identifiers for the unit reactions are collected in Table 4 for strands of up to six carbons, i.e., the 54 families defined above.

The composite refunctionalizations can now be generated as all combinations of two unit reactions having strands that overlap on at least one carbon. Once generated, they are all characterized by changing π-bonds at every bond in the overlapping strand (except of course for the simple redox composites on only one carbon, i.e., [R + R] and [X + X]). Hence they can all be divided into groups defined by their particular π-bond shifts. These characteristic π-bond shifts are listed in Table 5 in order of the strand length ($n = 1-6$), both for the unit reactions and the composites of two overlapped unit reactions. On the upright bonds in each structure are to be placed all combinations of H and Z changes, the numbers of which are also listed for each group at the left in Table 5.

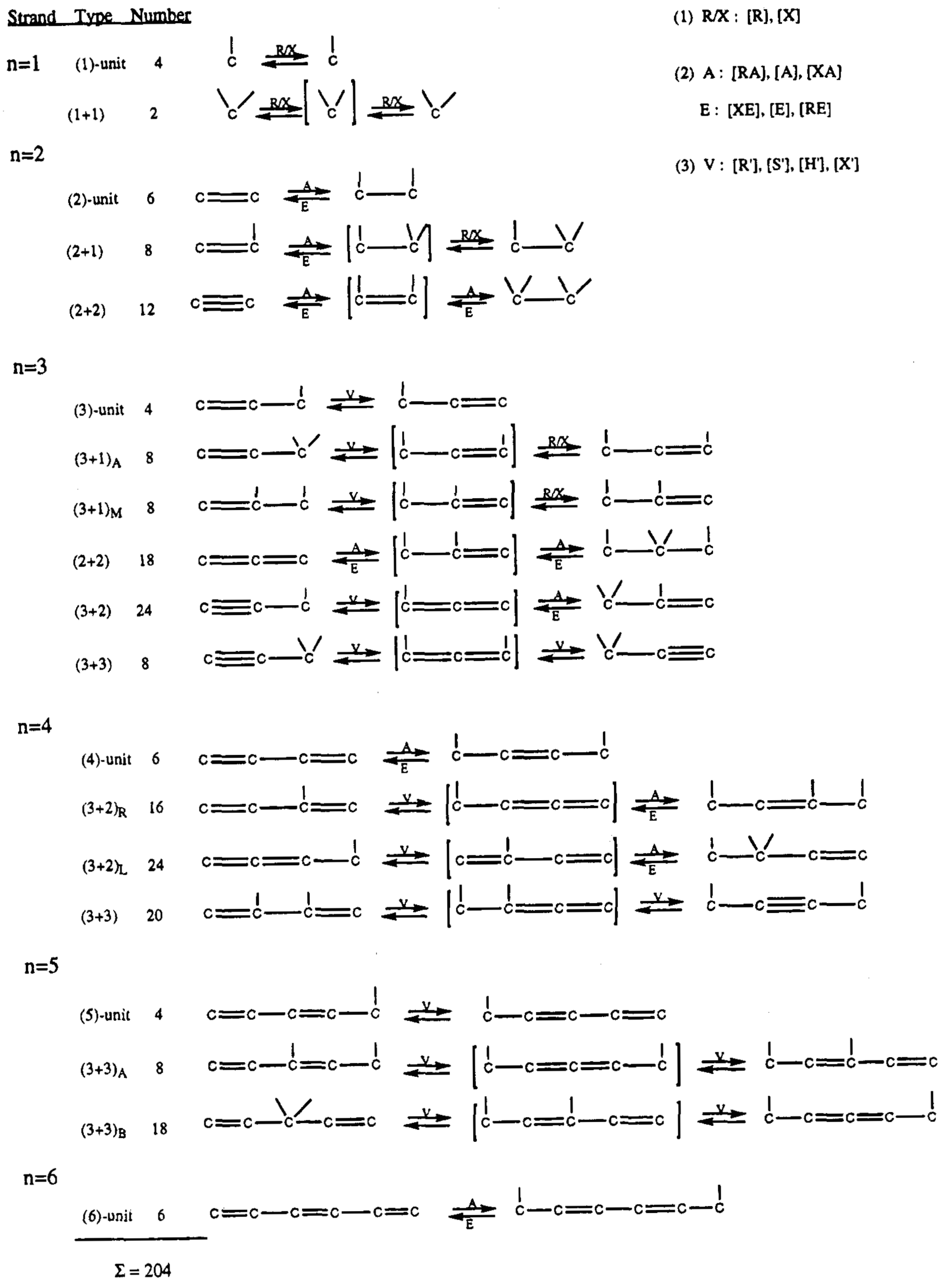
Table 5 shows all composites made from unit refunctionalization reactions with strands of 1−3 carbons, identified by their strand lengths, as with (2 + 1) for an addition/elimination of strand = 2 with an added redox on one of the carbons. This generation affords 174 composite families on strands up to five carbons, added to the 30 unit reaction families for refunctionalization. Longer strands with composites of further vinylogs can be constructed in the same way, but those shown in Table 5 are the set used by COGNOS and cover virtually every refunctionalization in several large databases.[5,9]

To illustrate the use of Table 5, the (3 + 2) composite on a three-carbon strand (24 cases) describes four possible vinylogous substitutions on a propargyl derivative to form the intermediate allenes in brackets, followed by one of the three addition reactions; this affords 12 forward and 12 reverse reactions. One of these would be [S′ + RA], characterized by a $\Delta z\pi$-list of **30E**; the reverse reaction [S′ + XE] is identified by **1ED**. If the two reactions of the composite are taken in reverse order, the overall change and the identification number are the same.

Two unit reactions on nonoverlapping carbons are treated as independent reactions in COGNOS, each identified and retrieved separately. These will be the cases in which one (or more) skeletal bond with unchanging π separates the reacting centers, as with the reduction C=C−C−Br → CH−CH−CH, annotated as the separate unit reactions [RA] and [R].

In a parallel manner we may create all the composites of single construction or fragmentation half-reactions with refunctionalizations which overlap on the same strand. The entry or loss of attachment H in Table 5 is paralleled by all the same forms with entry or loss of attachment R, since these do not alter the resultant $\Delta z\pi$-list. Thus the composites with skeletal change half-reactions can be derived in the same manner as the refunctionalizations from the forms in Table 5. Skeletal change half-reactions and refunctionalizations which differ only in replacing ±H with ±R will therefore

Table 5. Forms of Unit and Composite Reactions



often have the same $\Delta z\pi$-list, as may also be seen in the entries for Table 4, cf., **11** for [RA] and [RAC].

The COGNOS program affords a module to automatically read the entries in any given reaction database available in SYNLIB or REACCS format.[9] For each entry the program annotates the changing carbons of the reacting site into the σ, z, π, and h format and determines the $\Delta z\pi$-list. It then creates an index of these entries, arranged in numerical order of their family identification numbers within each skeletal class. This index groups all entries for each family together so that it becomes essentially instantaneous to retrieve the family which matches an input query and to display the number of entries found in it.

Refining the Match. In COGNOS any input query reaction is immediately assigned a $\Delta z\pi$-list, and the database entries matching that family are located in the index. In a large database the number of entries in any one of the major families will, however, be very large and unmanageable for searching. Therefore, a more refined definition of the query reaction (and the database entries) must be made to effectively reduce the number of matches found. Two main methods are advanced for this closer definition to prune down the number of hits in the retrieval operation.

Since the reaction class and family are already determined from the *changes* in skeleton and functionality ($\Delta\sigma$ and $\Delta z\pi$), the first refinement will be to compare the actual level of skeleton and functionality in the substrates. Therefore, we match the starting values of σ, z, and π at the changing carbons of the query against those recorded for the database reactions. For maximum flexibility in matching, COGNOS provides pruning keys for designating each of these values separately on the main carbons of the reacting site: in skeletal

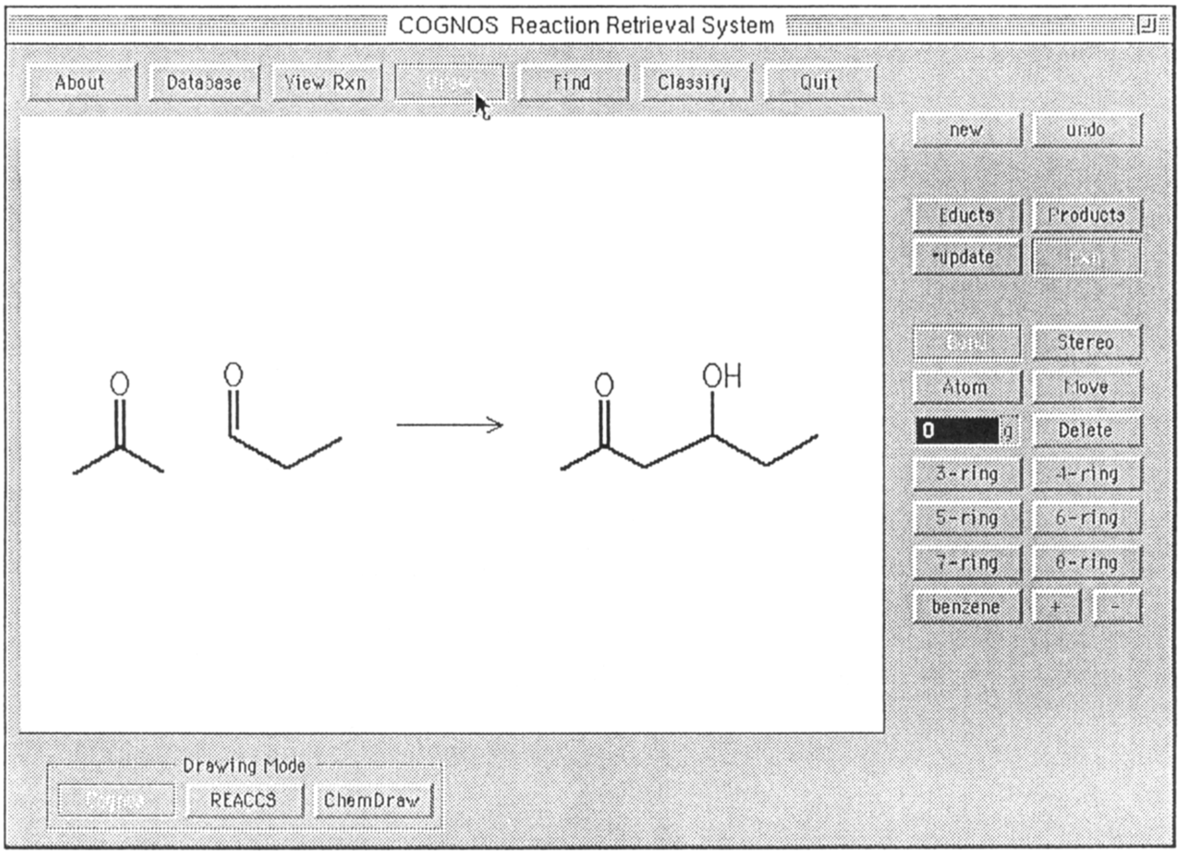


Figure 2. COGNOS drawing screen: query input.

changes only the ±R carbons themselves are compared; in refunctionalizations each end-carbon of the reacting strand is separately compared.

As an example, a Grignard reaction involves two half-reactions, each of which is matched. In the nucleophile half, if the reagent written in the query is EtMgX, with $\sigma = 1$ at the joining carbon, and then if this σ-value match is keyed, only those database entries with primary carbons ($\sigma = 1$) in the Grignard reagent will be matched and retrieved. In the electrophile half, the σ-value of the starting carbonyl may be matched ($\sigma = 2$ for ketones, $\sigma = 1$ for aldehydes or carboxylic derivatives) and/or the level of z may be used to match, as with $z = 2$ for aldehyde or $z = 3$ for carboxylic types ($z = 1$ would record the Grignard reacting with an epoxide or similar leaving group).

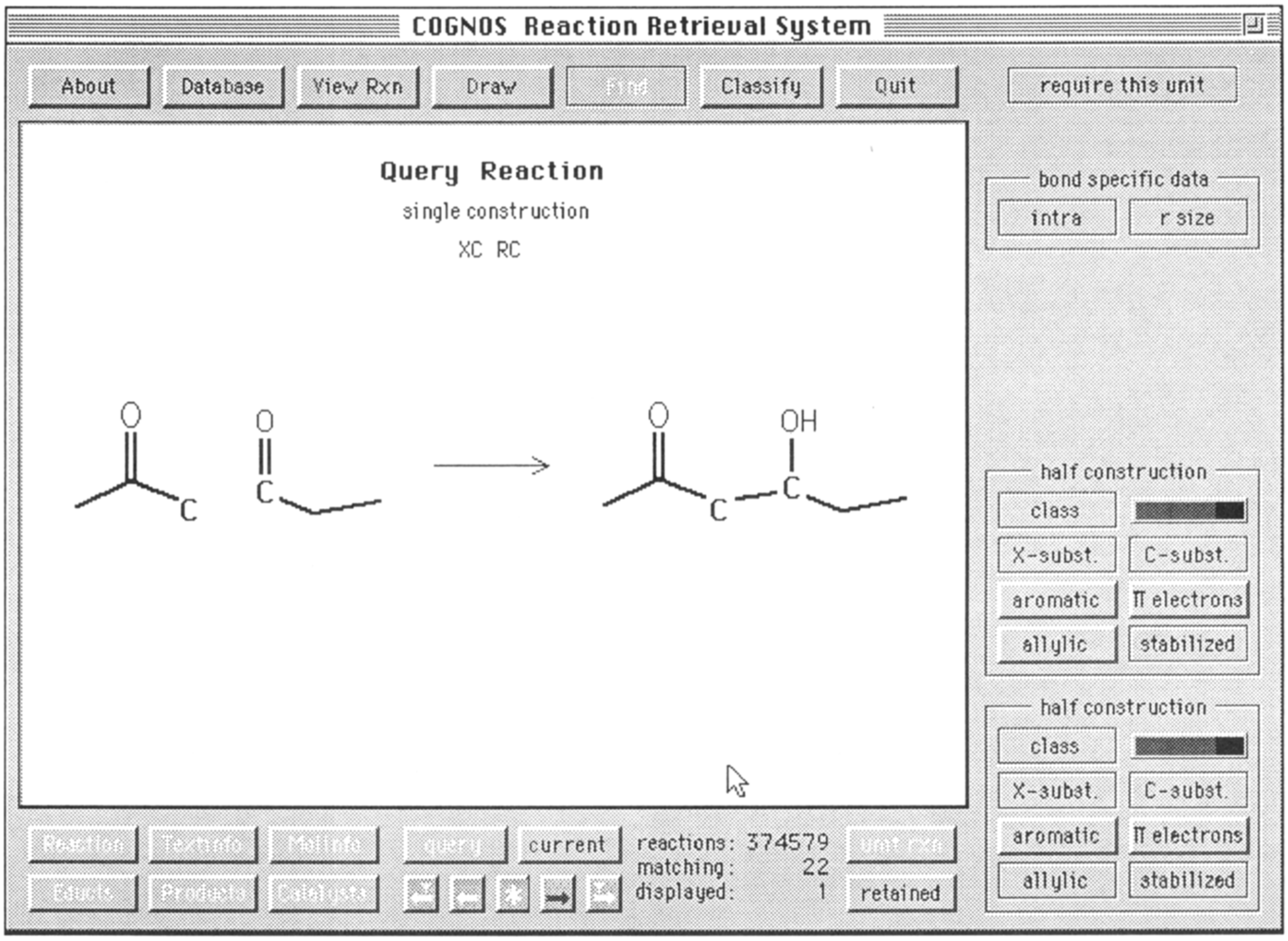
Matching the π-nature of any changing carbon as saturated, unsaturated, or aromatic is similarly available, as is the matching of any stabilizing or unsaturated group attached to it, and also the inter- or intramolecular nature of the reaction. For the most part, all these comparisons are simple matches of σ, π, z-values in the query with those of the database entries, and so are very rapidly made. Although these comparisons are still in the generalized σ, π, z-format, in most families the use of all the combinations of matching keys reduces the number of hits dramatically.

In the second matching mode, the detailed nature of any non-carbon atoms or groups lost from the substrate or gained in the product is compared. It is apparent in Tables 2 and 3 that in every unit reaction the end carbons of the strand will have +H, −H, +Z, or −Z, hence a gain or loss of either electropositive or electronegative atom. The procedure uses a binary search which allows for eight levels of increasing refinement in the comparison of the relevant atom type. For example, if the input query shows hydrogen lost from one carbon undergoing construction, one may match all entries with any kind of electropositive atom lost or successively refine the matching to various closer types of leaving atoms (cf., various metals, etc.) or just to hydrogen itself. Similarly, addition of bromine to a carbon atom may find all matches with any heteroatom, or more refined to any halogen, or to just bromine, with a decreasing number of hits after each closer refinement of the comparison.

In COGNOS the database entries are all indexed for these successive levels of matching. When a query is initiated, its class and family are generated first and the number of matching entries in its family ascertained. Within that family the program then follows a sequence of refinements, first in the σ, z, π-values of the substrates and then the more detailed nature of the entering and leaving groups. This is rapidly pursued by the program until it arrives at a number of hits deemed to be manageable for individual examination. The nature of the pruning choices made is then displayed on the pruning keys menu for the user, who may modify these himself, choosing either less or more refinement of matching and seeing displayed directly the number of hits which results.

A third matching mode available for the user identifies any functional group present and *not* changed in the reaction, allowing the user to specify if he wants only reactions which retain those groups unchanged in the product, as with a reduction at a reacting center which leaves a nitro group elsewhere unaffected.

Using the COGNOS Program. COGNOS is currently available for Macintosh computers on a simple CD-ROM disk coupled with an extensive database from InfoChem.[10] On entering the program the user enters the query reaction; a rapid and facile drawing mode is provided to enter just the reacting site of the query reaction. An alternative



Query Reaction

single construction

XC RC

reactions: 374579
matching: 22
displayed: 1

Figure 3. COGNOS search pruning screen: query reaction.

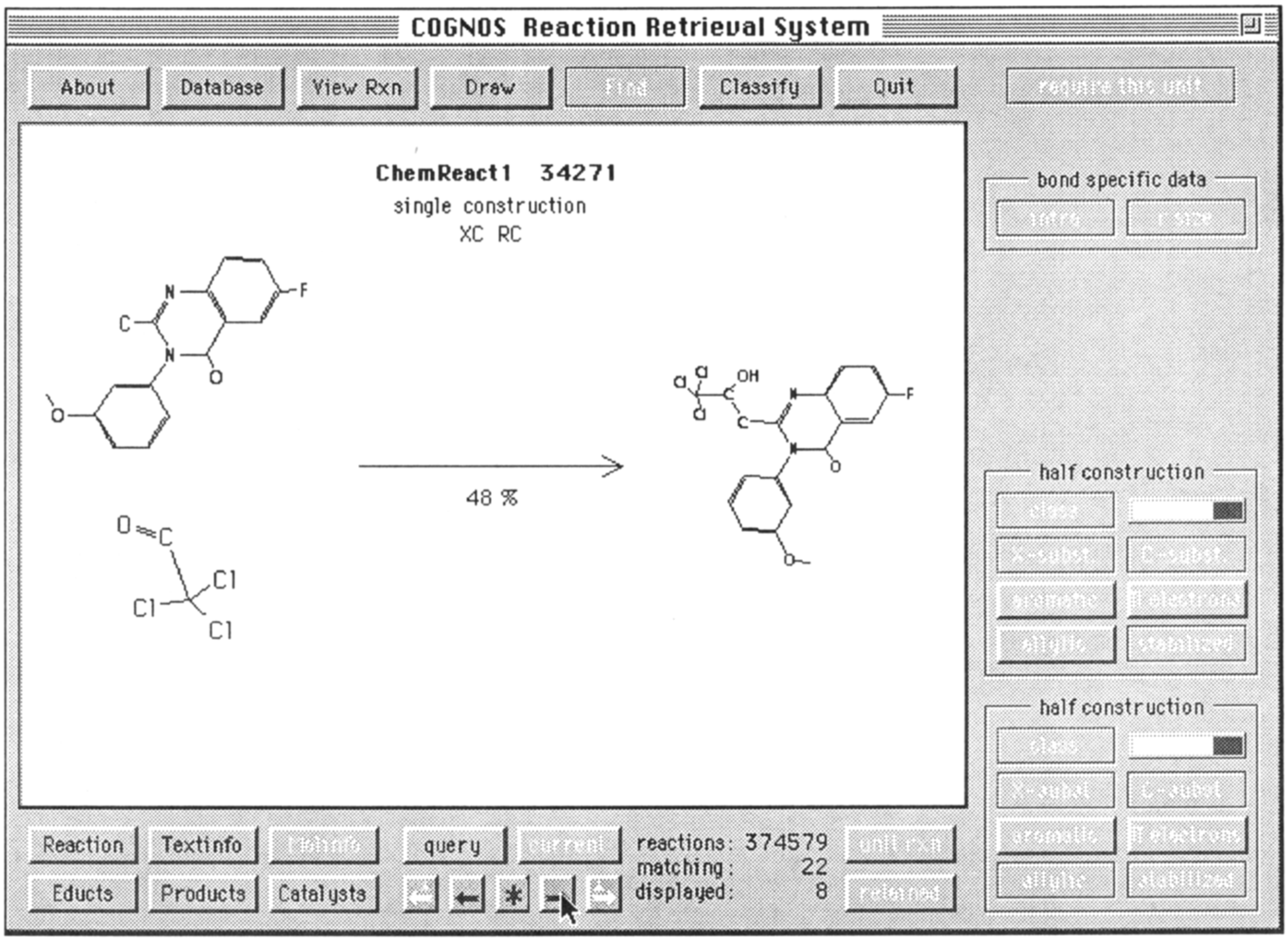
drawing mode using the ChemDraw format[11] is currently being implemented. The drawing screen is displayed in Figure 2, and each button may be activated by the mouse or the keyboard. The educt(s), i.e., substrate(s), may be drawn first with the **Educt** button down and then, when the **Product** button is pressed, the original structure(s) reappears ready for altering into the product(s); this ensures accurate atom mapping. The two may be entered in reverse order also. The whole query reaction may then be displayed with the **Rxn** button.

On activating the **Find** button the query reaction reappears (Figure 3) with a menu of keys for the refining and pruning operations described above. As the query reaction shown (aldol reaction) is a construction, a pruning menu for each half-reaction is displayed, and, as these buttons are pressed, the screen highlights the atom queried. Initially the **Find** button displays first the number of matching entries in the whole family and then goes quickly and automatically through a closer matching, sequentially showing the number of hits, until about 20 hits are found. The corresponding pruning buttons on the menu are automatically depressed as well so that these initial pruning choices may be seen. By pressing **Next** the user can then go directly to view these closest matching entries one at a time. He may at any time, by pressing **Query**, return to the pruning menu and change the level of refinement, seeing each time the number of hits in the current choice.

Besides the pruning buttons in each half-reaction box for the skeletal and functional nature of the changing carbons, there are the black matching bars used for refining the nature of heteroatoms entering or leaving in the reaction. These are activated by moving the cursor along the bar. Each time a refinement is made the screen shows the nature of the entering or leaving atoms selected with the bar, and the number of hits is revised immediately, as with the pruning buttons. Taken together, the pruning buttons and matching bars allow the number of matches to be manipulated until the user feels the selection is of manageable size. When a suitable number of matches is obtained, the matching entries themselves are then individually displayed, in order, using the **Next** button. Reaction entries found to be interesting may then be marked and/or printed out for inspection later.

The third pruning mode is represented by the **Retained** button, which brings up bars to indicate the functional groups present so that the user may then describe and refine any he wishes to see retained, unchanged, in the product. These matching bars operate like those above for the entering and leaving heteroatoms. An entry from the REACCS-CLF database[9] which matches the aldol query is shown in Figure 4. At bottom left of the menu are seen buttons to display the details of the reaction and the reference to the publication describing it.

Summary. The COGNOS program has been designed to meet the criteria described at the outset, generalizing all possible interconversions into unit (or composite) reaction families and indexing all the database entries by these families for instant retrieval. The system for doing this is fundamental in chemical terms and rigorous in application to digital format for the computer. The success of the program in finding correct matches from the databases lends



reactions: 374579
matching: 22
displayed: 8

Figure 4. Sample entry for input query.

confidence in this logic for organizing reactions, and the speed with which it finds matches from large databases makes it very practical to use. The literature precedents obtained are sharply defined by the system so that it is always clear just what results one will get and what will be excluded.

ACKNOWLEDGMENT

We wish to thank the Deutsche Forschungs Gemeinschaft for a postdoctoral fellowship for Dr. Sander. We also thank the Camille and Henry Dreyfus Foundation for partial support of this work through their Chemical Informatics Planning Grant.

REFERENCES AND NOTES

(1) Hendrickson, J. B. *J. Am. Chem. Soc.* **1971**, *93*, 6847; *J. Chem. Inf. Comput. Sci.* **1979,** *19*, 129.

(2) Hendrickson, J. B. *Angew. Chem., Intl. Ed. Engl.* **1990**, *29*, 1286.

(3) Hendrickson, J. B. *Recl. Trav. Chim. Pays-Bas* **1992**, *111*, 323.

(4) The term *isohypsic*, from the Greek for "equal level", was introduced to mean neither oxidative nor reductive.[1]

(5) Hendrickson, J. B.; Miller, T. M. *J. Chem. Inf. Comput. Sci.* **1990**, *30* , 403; *J. Am. Chem. Soc.* **1991**, *113*, 902.

(6) Hendrickson, J. B. *J. Chem. Educ.* **1985**, *62*, 245.

(7) The 32-bit word for an eight-carbon strand may be written in hexadecimal notation as 8 digits, and the changing carbons are placed at the right end of the word, leaving room at the left for the F-digits which will characterize negative $\Delta z\pi$-lists on any strand of less than eight carbons. In the discussion here the positive numbers are written without the left zeros and the negative numbers are underlined and written without the extra left F's. Thus a number **2FD** for an [S′] change at three carbons will be the word **000002FD**, and have a negative of $\underline{\textbf{D03}}$, to indicate **FFFFFD03**.

(8) Only two duplicated cases were found out of 313 identification numbers; these are the symmetrical composites **1E** and **22**; the first is either [A + RA] or [XE + E], and the second is either [RA + RA] or [E + E] with one E reversed, i.e. CHZ−CHZ → C≡C.

(9) The databases examined were as follows: SYNLIB from Distributed Chemical Graphics, Inc., Meadowbrook, PA.; and Theilheimer, Orgsyn, CLF and CSM in the REACCS format from Molecular Design, Ltd., San Leandro, CA.

(10) The major database from InfoChem GmbH, Munich, Germany, currently contains about 1.8 million reactions, of which some 370 000 are now present on the CD-ROM; more will be added later.

(11) *ChemDraw* is a product of Cambridge Scientific Computing, Inc., Cambridge, MA.

CI9401377