Estimate of Donor and Acceptor Sites Using Alternating Polarity Principle. Application to Pyridine Ring Construction

Dmitry E. Lushnikov

Institute of Organic Chemistry, Leninsky Prospect, 47, Moscow, 117913, Russia

Guido Sello*

Dipartimento di Chimica Organica e Industriale, Universita' degli Studi di Milano, via Venezian 21, Milano, Italy

Received February 13, 1995[⊗]

On the basis of the polarity alternation rule a quantitative model for the survey of the best donor and acceptor sites in polar organic molecules was developed. To this aim the following rules were suggested: (i) how to decrease induced polarity through the chain, (ii) how to evaluate the interaction between native and induced polarities; and (iii) how to take polar conflicts into account. The best donor and acceptor sites were assigned from literature data to the training set of 61 reagents used for pyridine and pyrrole syntheses (containing carbonyl, cyano, ester, and amino groups). The best correspondence between calculated and assigned ratings of best polar sites was determined using genetic algorithm. Four variables of the model and seven native bond polarities were included in the optimization. The fitness function was designed to reflect the degree of correspondence between calculated and experimental ratings. About 90% of compounds and 60% of reactions were correctly optimized by the obtained set of parameters.

INTRODUCTION

The perception of the best nucleophilic and electrophilic sites in a molecule is necessary for many applications including computer-assisted synthesis planning and reaction prediction. Donor and acceptor properties of atoms can be qualitatively evaluated by considering relationships between heteroatoms in organic molecules. The influence of heteroatoms in a carbon skeleton onto the reactivity can be accounted for by assigning charges to the carbon atoms in an alternating fashion.[1] In other words, the carbon atoms of a chain can be considered as having centers of donor and acceptor properties in alternating sequence. The first attempt to use the above polarity patterns for synthesis design was made in the TOSCA program.[2] Here the atoms of a molecule, starting from the heteroatoms, are characterized by plus or minus labels according to their acceptor or donor properties. Choosing the heteroatoms as the starting points of the analysis, minus−minus combinations coincide with heteroatom/heteroatom bonds. Other plus−plus and minus−minus combinations are located in the carbon skeleton as far away as possible from the heteroatoms. Molecules in which no two adjacent centers that have the same sign are called consonant.[2] Otherwise, they are dissonant. Accordingly, consonant reactions are charge-controlled reactions whose chemical driving force derives from the balancing of different charges of normal polarization. The preference of consonant reactions for the synthesis of consonant structures was postulated,[2] but for dissonant structures all reactions were deemed applicable.

In a more quantitative way the polarity alternation approach was used in the STRATOS program.[3] Here the resonance stabilization of charges was considered in order to decide which bond along the path between two functional groups should be disconnected. That bond will be disconnected that leads, when it is broken, to the best stabilization of the resulting charges. But only consonant paths were considered. Dissonant relationships need the "umpolung" of a donor or acceptor center,[4] and dissonant structures were not treated by STRATOS.

Recently we developed a program that carries out polar bonds disconnections and ranks them according to simple criteria.[5] Those disconnections where the direction of bond heterolysis coincides with polarity alternation pattern seems to be the most reasonable. For example, in the pyridine ring the most reasonable polarity pattern possesses the ring nitrogen negatively charged. Actually it was shown[6] that almost all one- and two-component syntheses of pyridine ring consistent with this pattern are known. On the contrary, there are only some known syntheses that are consistent with other, nonalternating, polarity patterns.

Unfortunately, consonant and dissonant relationships are only qualitative concepts. On the one hand, they consider only "heavy" heteroatoms, whilst hydrogen atoms are omitted. On the other hand, there were no attempts to develop rules that guide reactivity of dissonant molecules. For this reason we tried to develop the polarity alternation concept in a more quantitative way. In this paper we propose an algorithm for the calculation of nucleophilicity and electrophilicity indexes for atoms on the basis of polarity alternation patterns. The indexes are used to order the donor and acceptor sites in molecules and to select the best sites. The parameters of the algorithm were estimated for a particular class of reactions taken from pyridine chemistry. A genetic algorithm was employed to obtain the best correspondence between assigned and calculated ratings.

ALGORITHM FOR THE ESTIMATE OF DONOR AND ACCEPTOR SITES

The algorithm consists of four main steps.

[⊗] Abstract published in *Advance ACS Abstracts,* August 15, 1995.

0095-2338/95/1635-1060$09.00/0

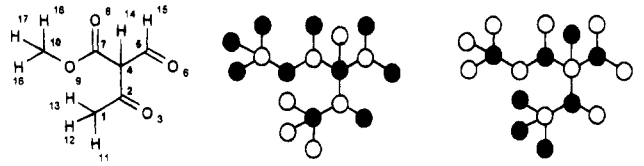


Figure 1. Possible patterns of polarity alternation for a sample acyclic compound. The two graphs represent complementary polarity alternation that can be used to assess the influence of different polarity assignements in determining the most stable polarity arrangement.

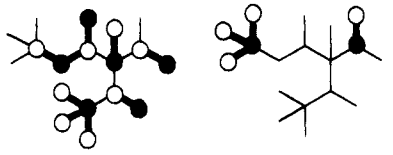


Figure 2. Native polar bonds (shown in **bold** face and in accordance to the corresponding polarity graphs) of the sample compound corresponding to different patterns of polarity alternation.

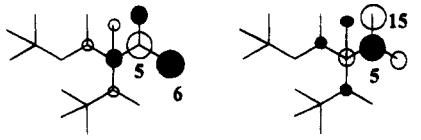


Figure 3. Patterns of induced polarities originated from the native polar bonds 5−6 (left) and 5−15 (right).

I. Assignment of Native Charges. We believe that each bond composed of atoms A and B of different kind is polar. As a consequence, it induces donor and acceptor centers on adjacent atoms, including hydrogens. Let each type of bond between atoms A and B be characterized by its own value of native polarity N_{A-B}. Then the most electronegative atom A becomes the donor center of intensity $-N_{A-B}$, and the least electronegative atom B becomes the acceptor center of intensity $+N_{A-B}$. In the following we will mark donor centers with heavy black dots and acceptor centers with empty circles. Each acyclic compound can be represented as a pair of bipartite graphs exhibiting two possible qualitative patterns of polarity alternation as illustrated in Figure 1.

Because C−C bonds are not polar, only the native polar bonds (i.e., those bonds whose polarity can be assigned by atomic electronegativity difference) are selected in Figure 2 and marked by bold lines.

II. Calculation of Induced Charges. Each native polar center induces on the adjacent atoms weaker donor and acceptor character in an alternating manner. At this step the donor and acceptor characters for each atom are stored separately. For example, bonds 5−6 (C=O) and 5−15 (C−H) in the sample structure produce the native and induced polar centers as shown in Figure 3.

It should be noted that the induced centers belong to the same bipartite graph as their parent centers. The reduction of circle size indicates a decrease of the strength of the corresponding donor/acceptor center. Given the polar bond A−B, the charge induced on atom C is obtained by multiplying the charge on atom B by a numeric coefficient. We use coefficient $\alpha_{sp}3$ if the pair of bonds A−B and B−C is nonconjugated, and coefficient $\alpha_{sp}2$ if the pair of bonds is conjugated. The induced charge on atom D is obtained by the same formula (Figure 4). In this case the coefficient is determined by the pair of bonds B−C and C−D. Cross-

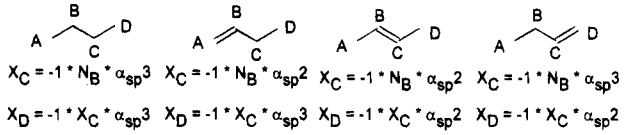


Figure 4. Calculation of induced charges on atoms. N, native polarity value; X, induced polarity value; and α is a coefficient ($0 < \alpha < 1$) depending on the type of bonds.

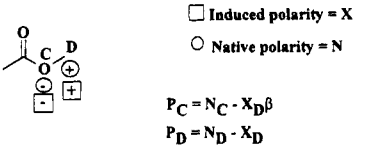
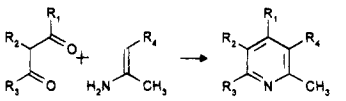


Figure 5. Calculation of atom polarity considering both native and induced polarities for atoms C and D.

Scheme 1



conjugation is not considered. Charges are induced only on the two nearest layers of connected atoms, because this approximation gives the best optimization results (see Discussion below).

III. Interaction of Induced Charges with Native Charges of the Same Sign. Two cases are distinguished: (a) An atom with an induced polarity value has no native polarity of the same sign. In this case the induced polarity is simply summed up with other induced polarities of the same sign. (b) Bond C−D (Figure 5) is polar, and induced polarities on atoms C and D have the same sign as their native polarities. In this case the induced polarity for the most distant atom D is subtracted from its native polarity (by absolute value). Let N_D be the native polarity of atom D, and X_D be the induced polarity of the same sign on this atom. Then the resulting polarity will be $N_D - X_D$. The polarity on atom C is also subtracted by $X_D * \beta$, where β is a parameter, $0 < \beta < 1$. The parameter should reflect the reduction of atom donor and acceptor properties caused by polarity alternation. It should be mentioned that at this step only polar effects of the same sign are taken into account. Thus the native polarity of an atom can be decreased up to zero, but it cannot change its sign.

IV. Taking Polar Conflicts into Account. After the influence of all polar bonds is calculated, it is necessary to consider centers with polar conflicts, i.e., that have both donor and acceptor characters. We suggest that the polar conflict leads to a small increase of both donor and acceptor properties. If D and A are the absolute values of the donor and acceptor indexes for the atom as calculated above, then we add $A*D*\gamma$ to both D and A values. Here γ is the polar conflict parameter, and $0 < \gamma \ll 1$. At last, we assign the ratings of donor and acceptor centers based on the calculated polarity indexes of all atoms.

SELECTION OF THE TRAINING SET FROM PYRIDINE SYNTHESES

To optimize the parameters of the algorithm it was necessary to build up a training set of compounds and to designate their best donor and acceptor sites. To this aim we selected a class of reactions where reaction sites can be directly deduced from the evaluation of reaction products,

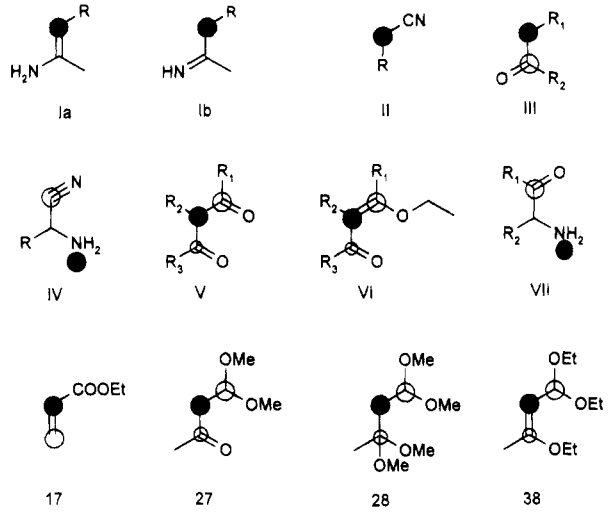


Figure 6. Types of compounds included in the training set. Substituents are listed in Tables 1−7.

Table 1. Compounds of Type **I**

no.		
a	b	R
1	2	CN
3	4	COMe
5	6	COOEt
7	8	COPh

Table 2. Compounds of Type **II**

no.	R
9	$CONH_2$
10	CN
11	COOEt

Table 3. Compounds of Type **III**

no.	R_1	R_2
12	COOEt	COMe
13	COOEt	Ph
14	Ph	Me
15	Ph	Ph
16	CH_2COOEt	COOEt

avoiding the evaluation of reaction network or of multiple intermediates. We found that these conditions are satisfied by reactions of 1,3-bielectrophiles (1,3-dicarbonyl compounds and their synthetic equivalents) with 1,3-binucleophiles (enamines) leading to pyridines.

Usually the condensations with enamines occur in mild conditions without adding basic or acidic catalysts, while classical Guareschi synthesis, that uses cyanoacetamide (compound VI, R = $CONH_2$ in Figure 6) instead of enamine, requires the addition of a base. It is assumed[7] that in the first step the most active carbonyl group of the 1,3-diketone reacts with the most nucleophilic methylene group of the enamine. Then another carbonyl group reacts with the amino group closing the cycle. We were only interested in nonsymmetrical dicarbonyl compounds ($R_1 \neq R_3$). In this case the most active carbonyl group will be positioned at the R_1 substituent.

All 61 compounds included in the training set are given in Figure 6; the substituents are listed in Tables 1−7. In Figure 6 the ratings of centers used for the optimization are shown. Bigger empty circles mark the best acceptor centers, and smaller circles (if present) show centers of successive

Table 4. Compounds of Type **IV**

no.	R
18	H
19	Me
20	Ph
21	COOEt

Table 5. Compounds of Type **V**

no.	R_1	R_2	R_3
22	H	$-(CH_2)_4-$	
23	H	H	Ph
24	H	COOEt	Me
25	COOEt	H	Ph
26	COOEt	H	Me
29	H	Me	Me
30	H	Me	Ph
31	Me	H	Ph
32	$CONH_2$	H	Me
33	CH_2OMe	H	Me
34	CH_2OMe	OCOMe	Me
35	Me	H	OEt
36	$-(CH2)_4-$		OEt
37	COOEt	H	OEt
39	CH_2COOEt	H	OEt
40	H	COOEt	COOEt

Table 6. Compounds of Type **VI**

no.	R_1	R_2	R_3
41	H	COMe	COOEt
42	H	COOEt	COOEt
43	H	COOEt	Me
44	Me	H	Ph
45	Me	H	COOEt
46	H	COOEt	OEt
47	H	Ph	OEt

Table 7. Compounds of Type **VII**

no.	R_1	R_2
48	OEt	H
49	OEt	Me
50	Me	H
51	H	H
52	Ph	H
53	Me	Me
54	H	Me
55	$CH_2CH_2NH_2$	H
56	Ph	Ph
57	Ph	COOEt
58	Me	COOEt
59	CH_2COOEt	COOEt
60	CH_2CH_2COOEt	COOEt
61	Me	COMe

rating. Donor centers are marked by black circles. For example, for 1,3-dicarbonyl compounds of type V the strongest acceptor center is at R_1 and the second is at R_3, while the strongest donor center in the molecule is located between them. The structures of 1,3-bielectrophiles (V, VI, **27, 28,** and **38**) and of 1,3-binucleophiles (I and II) were taken from a review on pyridine syntheses.[7]

As it is not known what tautomeric forms actually react, we accepted the diketo form for the dicarbonyl compounds of type V, and we also included several enol-ethers of type VI. However, for the enamines I we selected for the optimization both imine and enamine tautomeric forms.

It must be noted that the best donor site in the 1,3-bielectrophile should not be better than that in the binucleophile. In addition the best acceptor atom in the 1,3-

Table 8. Pairs of Compounds Used for the Reaction Optimization and Results of Optimization

	1,3-binucleophile		
1,3-bielectrophile	both are correct	donor is incorrect	both are incorrect
22	2, 4, 6	1, 3, 5, 9	
23	3, 4, 5, 6, 9		
24	6	5	
25	3, 4, 5, 6, 9		
26	5, 6, 8, 9	7	
27	9		
28	9		
29	4	3, 9	
30		9	
31	9, 10		
32	9		
33	9, 10, 11		
34		9	
35	9		
36			9
37	9		
38	1, 2, 3, 4, 5, 6, 7, 8, 9		
39	9		
41[a]			1, 2, 3, 4, 5, 6
42[a]			1, 2, 3, 4, 5, 6
43	6	5	
44	9		
45[a]			9
46	6	5	
47			9

[a] Reactions are incorrect because order of two best electrophilic centers in bielectrophiles is reversed.

dicarbonyl compound should be more powerful than any acceptor in the binucleophile. We included in the training set 69 reactions between our compounds taken from the same review.[7] All the reactions summarized in Table 8 are simply pairs of compounds (electrophile−nucleophile) where the above mentioned "reaction" relationships should be satisfied. We must emphasize that the training set is actually made by two subsets: one containing the reacting compounds (both electrophiles and nucleophiles); the other containing the simulated reactions between members of the first subset in the sense just described.

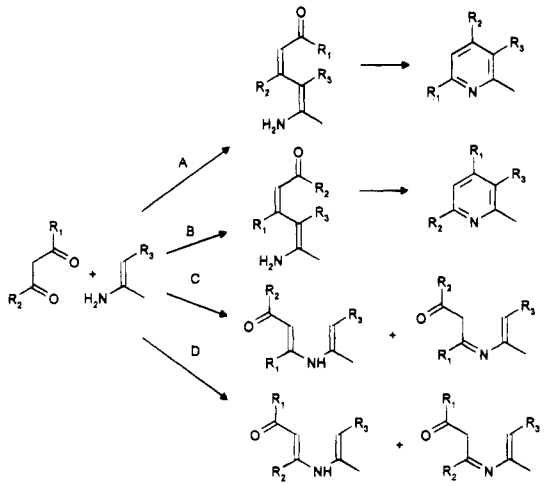
More, we included in the compound training set some -aminoketones (VII) and -amino nitriles of type IV taken from another review.[8] Their condensations with -methylene ketones (III) are well-known as Knorr pyrrole synthesis. In this case only the best donor and acceptor centers in the molecules are given, because there are no definite rules to deduce what pair of centers should react first. For this reason reactions using these compounds were not included in the training set.

SIMULATION OF REACTIONS BY CAMEO PROGRAM

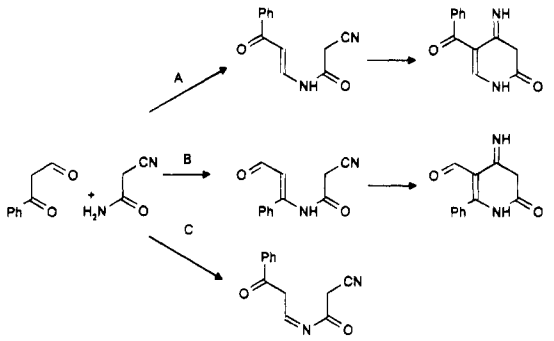
It was interesting to investigate intermediates and products of reactions between enamines and 1,3-dicarbonyl compounds predicted by the CAMEO program.[9] The predicted reaction pathways are considerably different for the enamines I and for the cyanoacetamide used as binucleophiles.

(a) Reactions of 1,3-Dicarbonyl Compounds with Methylenic Enamines. In this case the program CAMEO uses the heterocyclic module. Without addition of basic or acidic reagents in the temperature range of 50−100 °C, the program smoothly predicts the formation of the appropriate pyridines (Scheme 2). For the case of simple reagents (R_1 = H, R_2 = Ph, R_3 = CN, COCH3) the only reaction is the condensation of carbonyl group with the methylene group of the enamine. In the second step another carbonyl group reacts with the amino group leading to the pyridine ring (pathways A and B in Scheme 2).

Scheme 2



Scheme 3



The enthalpy of reaction calculated by CAMEO indicates the preferential reaction of the most active carbonyl group. However, the enthalpy of two-step reactions is identical for both pyridine isomers.

In the case of R_1 or R_3 = COOEt the initial reaction of the amino group of the enamine is also predicted by CAMEO (Scheme 2, pathways C and D) (the resulting imines were found experimentally in some cases), but they cannot be transformed into the title pyridines, according to CAMEO. If there are other possibilities for cyclization (for example, other cyano or ester groups are present), the program almost always predicts cyclizations involving these groups.

Quite different pathways are predicted for the reaction of cyanoacetamide with 1,3-dicarbonyl compounds (Scheme 3). This reaction is catalyzed by bases and requires higher temperatures. The CAMEO program predicts that cyanoacetamide will react with dicarbonyl compound only by the amino group and only in the presence of a base like NaOEt. The resulting intermediates under no circumstances form target pyridines.

Even if one specially deprotonates cyanoacetamide to form the anion at the methylene group, CAMEO uses the nucleophilic rather than the heterocyclic module. Among plenty of predicted intermediates the required pyridines were not present. We succeeded in obtaining the product of correct ring closure. However, it was not possible to predict its dehydration and aromatization to the title pyridone.

As a general result, the use of the CAMEO program allowed us to confirm the mechanism of pyridine ring formation and to validate ascribed ratings of donor and acceptor centers. At the same time, CAMEO cannot distinguish two possible pyridine regioisomers. Reactions with cyanoacetamide seem to be more complex and may be in only formal agreement with our ranking of donor centers in cyanoacetamide.

APPLICATION OF GENETIC ALGORITHM TO PARAMETER OPTIMIZATION

To find the parameter values suitable for the correct prediction of the best donor and acceptor sites in the training set, a genetic algorithm (GA) was employed. GA is known to be a robust optimization technique useful for finding global optima or suboptima.[10] GA operates with one-dimensional arrays of numbers usually called chromosomes. In our case the chromosome consists of floating-point numbers representing the parameters of the algorithm. Given the chromosome, a fitness function calculates the ratings of all donor and acceptor sites in the molecules of the training set and compares them with assigned ratings. The resulting penalty value reflects the quality of the given set of parameters in respect of making reasonable predictions. The aim of the genetic algorithm is to find sets of parameters with minimum penalty values. The genetic algorithm itself starts from random values of all parameters (within given ranges) and finds the suboptimal sets of parameters. We used the GA package GENET available through INTERNET.[10] This package can operate with chromosomes made by real numbers. The user should only set up the range for each parameter to be optimized and supply the fitness function.

The penalty for each molecule is the sum of the penalties for all the atoms with ascribed ratings (i.e., atoms marked on Figure 6). Let the atom i possess the rating n (n = 1,2), but for the given chromosome its rating is calculated to be m, while the rating n is given to another atom j. Let also $V[i]$ be the calculated polarity index for the atom i, while $V[j]$ is the corresponding value for the atom j. Then the penalty for atom i is calculated as follows

$$P(i) = 100*\mathrm{Abs}(n - m)*[\mathrm{Abs}(V[i] - V[j]) + 1]/n$$

The multiplier (100) is arbitrarily chosen to get penalty values grater than 10. The penalty is increased with increasing the error in rating determination. It is expressed by the $\mathrm{Abs}(n - m)$ multiplier. The penalty will decrease for sites with minor ratings (2 and more). This is achieved by the division by n. The difference in absolute values $\mathrm{Abs}(V[i] - V[j])$ should help GA to reduce more quickly the error in rating calculation.

In order to improve the understanding of the calculation of the penalty values we will report an example of calculation. Let us consider the structure V of Figure 6 (with R_1 = CH_3, R_2 = H, R_3 = OEt) and let us suppose that the best electrophilic center is the ketone carbon atom 1; thus the assigned ratings are respectively position 1 for atom 1 and position 2 for the carboxy carbon atom 2. The two atoms have the polarity indexes 0.9 and 0.73. In the case of correct assignement by the GA the penalty value is equal to 0. On the contrary, if the calculated ratings are incorrect, then the penalty is calculated as follows

$$P(1) = 100*\mathrm{Abs}(1-2)*[\mathrm{Abs}(0.9 - 0.73) + 1]/1 = 117$$

$$P(2) = 100*\mathrm{Abs}(2-1)*[\mathrm{Abs}(0.73 - 0.9) + 1]/2 = 58.5$$

and the total penalty will be equal to 175.5.

If all donor (or acceptor) centers are predicted correctly, the penalty from the above formula is equal to zero. If there are two or more atoms with assigned ratings of the same polarity, an extra penalty is calculated as

$$P(i,j) = 1/\mathrm{Abs}(V[i] - V[j])$$

where $V[i]$ and $V[j]$ are absolute values for the two centers with ratings n and $n + 1$. This small addition plays its role when the ratings are correct, but the gap between the corresponding absolute values is very small.

Another important part of the fitness function is the reaction penalty. The reactions are simply represented by pairs of compounds (**E**, **N**), where the former is the electrophile and the latter is the nucleophile. Hence the best donor center in the molecule **N** should be even more nucleophilic than the best donor center in the molecule **E**. Vice versa, the best acceptor center in the molecule **E** should be even stronger than the best acceptor center in the molecule **N**. If in one of the molecules the best donor or acceptor site is predicted incorrectly, the penalty for this reaction is 2*30, i.e., 60. If centers are predicted correctly but one or both reaction rules are violated, the penalty for each violated rule is calculated as

$$P(a,b) = 30*[\mathrm{Abs}(V[a] - V[b]) + 1]$$

where $V[a]$ and $V[b]$ are values of donor (acceptor) strength for atoms with rating equal to 1 in both molecules. If the rules are fulfilled, the penalty of the reaction is equal to zero. Finally, to obtain the penalty value for the whole training set, all molecule and reaction penalties are summed up.

RESULTS

Parameter Optimization. The training set contained seven types of bonds (Table 9). The ranges of all native bond polarities were arbitrarily set from 0 to 10. Because only the ratio of native bond polarities is essential, after several runs we fixed the polarity of H−C bond to be 3.0. The ranges for parameters α_{sp^3}, α_{sp^2}, β, and γ were set from 0 to 1.

For these parameter ranges the optimized minimal value of the fitness function strongly depends on the population size and slightly depends on the number of generations. Hence after 5−10 generations the GA falls into some local minima, whilst it is necessary the use of very large population size (10.000 chromosomes) to reach the actual minimum of the fitness function. This fact may be caused by the very broad initial ranges of parameters.

To avoid time-consuming calculations using large pool size, we developed a program shell for the dynamic narrowing of the parameter ranges using pool size as small as 10 chromosomes. The shell maintains a list of the six best chromosomes through several runs. Currently, it runs GA six times, selects the three best runs with minimal penalties, and puts these best chromosomes into the list by replacing the poorest chromosomes. Then the lowest and the highest values of each parameter from the three current best chromosomes are selected to be the new parameter range.

Table 9. Statistics of the Training Set

type of bond	H−C	H−N	C−O	C=O	C−N	C=N	C≡N
no. of bonds	593	54	110	97	25	4	10

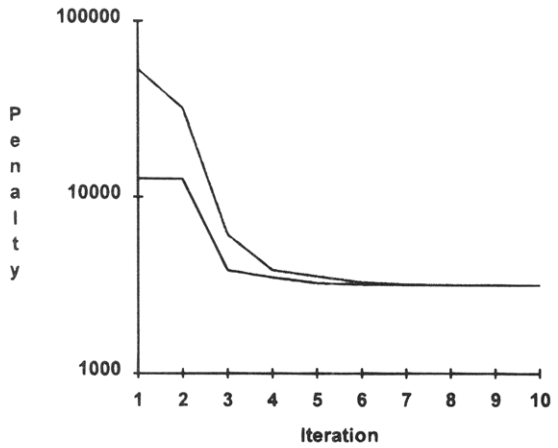


Figure 7. Change of the penalty range after several recalculations of parameter ranges at 10−chromosome pool.

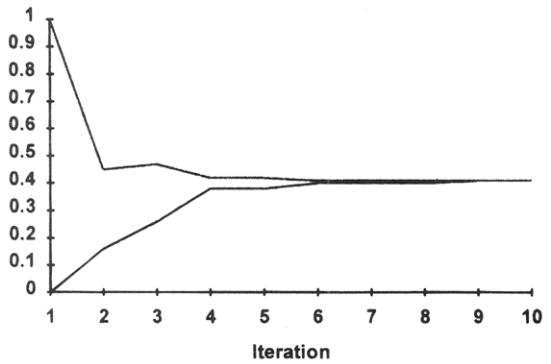


Figure 8. β Parameter ranges recalculated from the best chromosomes at 10-chromosomes pool.

Finally, the bounds of the new ranges are expanded by 20% to the top and to the bottom. This procedure (somewhat similar to GA selection) is repeated up to 10 times. The changes of the penalty range over several iterations are presented in Figure 7; the changes for β parameter are shown in Figure 8.

The range recalculation allowed us to sufficiently reduce the calculation time (from several hours per run to 30 min on an IBM PC AT/486, 33 MHz) with approximately the same resulting penalties and parameter values.

The ranges of the four algorithm parameters obtained at pool size 10 and pool size 10.000 are presented in Figure 9, and the corresponding values of the bond polarities are given in Figure 10.

Prediction of the Best Polar Sites in the Training Set. Using pool size of 10.000 chromosomes the correct ratings were calculated for 56 compounds on 61. In Figure 11 the structures with incorrectly predicted ratings of acceptor sites are shown. For compounds **40, 41, 42,** and **45,** the best two electrophilic centers were predicted correctly, but their order was inverted. For compound **37** the best electrophilic center was predicted correctly, but the second and third electrophilic center were inverted. (The third centre is the COOEt group not marked by an empty circle).

Prediction of the Reactions in the Training Set. In 69 reactions 42 were correctly optimized, 13 were not optimized

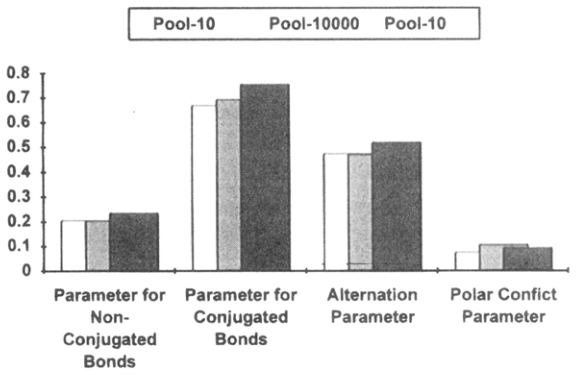


Figure 9. Range of parameters obtained at pool size 10 and at pool size 10.000. The first and the third columns are the range limits obtained using pool size equal 10. Parameters are as follows: α_{sp^3} for nonconjugated bonds (first column: 0−0.206; second column, 0−0.206; third column, 0−0.236), α_{sp^2} for conjugated bonds (0−0.67, 0−0.694, 0−0.755), β for alternation parameters (0−0.474, 0−0.47, 0−0.52), and γ for polar conflict parameters (0−0.075, 0−0.106, 0−0.093).

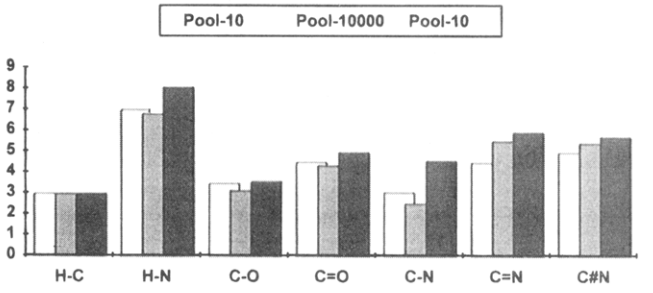


Figure 10. Range of bond polarities obtained at pool size 10 and at pool size 10.000. The first and the third columns are the range limits obtained using pool size equal 10; H−C (first column, 0−3; second column, 0−3; third column, 0−3); H−N (0−7.027, 0−6.82, 0−8.076); C−O (0−3.486, 0−3.14, 0−3.582); C=O (0−4.509, 0−4.34, 0−4.977); C−N (0−3.051, 0−2.52, 0−4.574); C=N (0−4.471, 0−5.52, 0−5.927); C#N designates triple bond in cyano group (0−4.969, 0−5.42, 0−5.723).

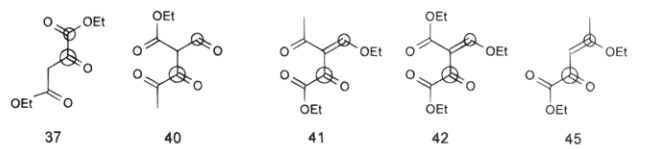


Figure 11. Structures with incorrectly predicted best acceptor sites. The predicted order of sites is shown. For compounds **40, 41, 42,** and **45,** the order of two best sites is inverted, for compound **37** the second site should be at the another COOEt group.

due to wrong prediction of best acceptor center, in 12 reactions the best donor center in the binucleophile was weaker than in the bielectrophile, and in two reactions both reaction rules were not satisfied, as summarized in Table 8. To display incorrect reactions we represented in Figure 12 the relation between the best donor and acceptor indexes for all compounds included in Table 8.

DISCUSSION

The proposed approach to the estimate of the best donor and acceptor sites in a molecule is based on the idea of polarity alternation, supplemented by several assumptions. First, polar influence of heteroatoms decreases through the chain, and it is well-known that it decreases more quickly for saturated chains than for conjugated chains. We selected the maximal length of a chain to be three bonds starting from

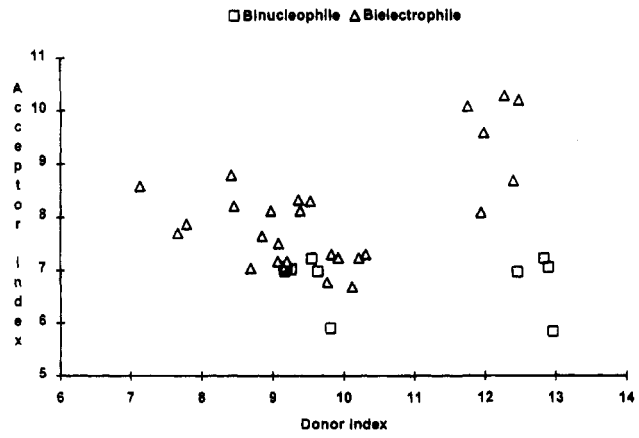


Figure 12. Plot of best acceptor vs best donor indices for compounds **1**−**11** and **22**−**47** used for reaction optimization.

the heteroatom. The first is the C−Heteroatom bond, and the other two bonds may be of any type. Longer chains give worse optimization results. Our second assumption concerns the relationship between the strength of the alternation and the nucleo- or electrophilicity of atoms. We believe that the induced polarity affects polar and nonpolar bonds differently. The induction of polarity on a nonpolar bond (C−C) makes it more polar and more sensitive to heterolysis and to other polar reactions. Making a polar bond (for example, C=O) more polar makes it more stable and less sensitive to nucleophilic addition. Therefore the ability of the carbonyl carbon to add nucleophiles should increase in the sequence O=C=O < O−C=O < C=O. Similarly, the acetylation of amines decreases the nucleophilicity of the nitrogen atom due to appearance of a consonant amide fragment N-C=O. The third assumption accounts for the nucleo-and electrophilicity of conflict atoms, i.e., atoms that have both nucleophilic and electrophilic properties. Our solution was inspired by the fact that the electrophilicity of a carbonyl group is decreased in the order CH_2=O > C−CH=O > C−C(=O)−C. Thus the higher electrophilicity of aldehydes can be related to a polar conflict in formaldehyde and other aldehydes. It is also known that α-dicarbonyl compounds are more electrophilic than simple ketones. This can also be explained by the influence of polar conflict. In our scheme the presence of a polar conflict slightly increases both the nucleophilic and the electrophilic characters of the given atom.

Altogether we introduced four parameters: two parameters reflect the decrease of polarity through the chain (for conjugated and nonconjugated bonds), and then there is one parameter for consonant fragments and one parameter for dissonant fragments. Other variables concern the native bond polarities for all types of bonds present in the training set.

Using the genetic algorithm (that starts from random values) we were able to find the values of all parameters that give correct prediction of the best donor and acceptor sites for 90% of the training set. The native bond polarities obtained (Figure 10) are in good correspondence with the order of atom electronegativities. For example, the polarity of single bonds increases in the order C−H < C−N < C−O; the polarity of H−N bond is approximately equal to the sum of polarities of H−C and C−N bonds. The polarities of multiple bonds are greater than the polarities of the corresponding single bonds. Because we have only four compounds with C=N bonds, its native polarity was estimated very roughly and in some runs exceeded the polarity of C≡N bond.

The parameters of the algorithm vary in the range 0−1 and therefore can be expressed in percents. Only 20% of the polarity is induced to the next atom in the nonconjugated chain, while in the conjugated chain the transfer is about 70% of the starting value. Actually, polarity effects are much more important for conjugated compounds. The value of the polar conflict parameter was found to be about 10%. It indicates that the polar conflicts actually can lead to the increase of both donor and acceptor properties of atoms. This effect is small but not negligible. The value of the alternation parameter (50%) points out that the philicity of the central atom in consonant fragments should sufficiently decrease due to the polarity alternation.

It should be noted that the major part of the training set (taken from pyridine chemistry) is represented by structures with polarity alternation in the main chain (1,3-dicarbonyl compounds, enamines). In all the compounds the two best acceptor sites and the best donor site were correctly located, except for the five structurally related compounds shown in Figure 11. The errors for these compounds are not very serious: the two best acceptor centers were exchanged. These compounds clearly demonstrate that the influence of polar conflicts on the carbonyl group, caused by the COOR group, was overestimated. Nethertheless, α-aminonitriles and α-aminoketones (possessing polar conflict between the best donor and acceptor sites) were optimized correctly.

The relations between the best donor and acceptor centers formulated for the reactions were introduced to prevent GA from making some centers very nucleophilic or electrophilic. In spite of considering an ideal situation (no influence of base, media, etc.) about 60% of reactions were optimized correctly. Let us consider the values of donor and acceptor indexes as shown in Figure 12. Both bielectrophilic and binucleophilic reagents form two clusters. Bielectrophiles containing the electron-withdrawing −COR group at their donor center (**24, 40, 41, 42, 43,** and **46**) form the cluster with higher donor and acceptor indexes, situated at the top right corner of Figure 12. The binucleophiles of the type Ib (imines) possess high donor index at the methylene group and form the cluster at the right bottom corner of Figure 12. Binucleophiles in the enamine form (**1, 3, 5,** and **7**) and compounds of type II (**9**−**11**) are characterized by much lower donor index value. Hence, pairs of compounds where the binucleophile center is situated to the right and lower of the bielectrophile center correspond to the correctly predicted reactions.

CONCLUSION

The polarity alternation offers an alternative way of thinking to reactivity problems as well as a simple basis for the computer-assisted reactivity modeling. However, the proposed scheme does not represent a comprehensive and complete approach to the estimate of reaction sites in polar molecules. It was developed to point out that the polarity alternation can be transformed from a mnemonic qualitative rule to a quantitative reactivity concept. In the present work we proposed a simple way to do it. The most important assumption concerns how to take dissonant relationships into account. It allowed us to obtain good results in the particular field of pyridine syntheses. However, we selected the

simplest formulas to calculate polar influence and to take polar conflicts into account. Therefore, they may not be applicable in all cases. Another problem is that the nucleo- and electrophilicity actually depends on the other reaction partner. In the dissonant molecules (bromoacetone, for example), the same atom (α-carbon between carbonyl and bromine) can act either as an electrophile or as a nucleophile, depending on the reaction. These problems will be addressed in our future studies.

ACKNOWLEDGMENT

We are deeply grateful to the CNR (Progetto Finalizzato Chimica Fine II), Rome for a grant and to Professor N. S. Zefirov (Moscow State University) for making possible this research. Partial financial support (to D.L.) from International Science Foundation under Grant M1X000 and from Chemical Structure Association Trust is gratefully acknowledged. We also express thanks to Dr. E. Babaev (Moscow State University) for valuable discussions. The author gratefully acknowledges the possibility of using CAMEO which has been made available by Prof. W. L. Jorgensen; sincere thanks also goes to Prof. Carlo Scolastico.

REFERENCES AND NOTES

(1) Ho, T.-L. *Polarity Control for Synthesis*; Wiley: Chichester, 1991.

(2) Doenges, R.; Groebel, B. T.; Nickelsen, H.; Sander, J. TOSCA: A Topological Synthesis Design by Computer Application. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 425−430.

(3) Wagener, M.; Gasteiger, J. Implementation of Synthesis Strategies in Prolog. In *Software Development in Chemistry 4*; Gasteiger, J., Ed.; Springer-Verlag: 1990; pp 265−273.

(4) Seebach, D. Methods of Reactivity Umpolung. *Angew. Chem., Int. Ed. Engl.* **1979**, *18*, 239−258.

(5) Lushnikov, D. E.; Babaev, E. V. Molecular Design of Heterocycles. 4. Application of computers in heterocyclic chemistry. *Khim. Geterotsikl. Soedin.* **1993**(10), 1299−1318 (Russ.).

(6) Babaev, E. V. Molecular Design of Heterocycles. 2. Magic "Structure-Synthesis" Rules for 6-Membered Heteroaromatic Rings Synthesis. *Khim. Geterotsikl. Soedin.* **1993**(7), 937−961 (Russ.).

(7) Brody, F.; Ruby, P. R. Synthesis and Natural Sources of the Pyridine Ring. In *The Chemistry of Heterocyclic Compounds. Pyridine and Its Derivatives. Part I*; Klingsberg, E., Ed.; Wiley Interscience: 1960.

(8) Bean, G. P. The Synthesis of 1H-Pyrroles. In *The Chemistry of Heterocyclic Compounds. Pyrroles. Part I*; Jones, R. A., Ed.; Wiley Interscience; 1990.

(9) Jorgensen, W. L.; Laird, E. R.; Gushurst, A. J.; Fleischer, J. M.; Gothe, S. A.; Helson, H. E.; Paderes, G. D.; Sinclair, S. CAMEO: A Program for the Logical Prediction of the Products of Organic Reactions. *Pure Appl. Chem.* **1990**, *62*, 1921−1932.

(10) Goldberg, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning*; Addision-Wesley; Reading, MA, 1989.

(11) GenET version 1.0 (author C. Z. Janikow) is available by anonymous FTP at **radom.umsl.edu** in the file **/var/ftp/GenET.tar.Z**.

CI9502010