(3) Herndon, W. C. On Enumeration and Classification of Condensed Polycyclic Benzenoid Aromatic Hydrocarbons. *J. Am. Chem. Soc.* **1990**, *112*, 4546-4547.
(4) Dias, J. R. A Periodic Table for Polycyclic Aromatic Hydrocarbons. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 139.
(5) Dias, J. R. Studies in Deciphering the Information Content of Chemical Formulas: A Comprehensive Study of Fluorenes and Fluoranthenes. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 2-11.
(6) Schmidt, W.; Grimmer, G.; Jacob, J.; Dettbarn, G.; Naujack, K. Polycyclic Aromatic Hydrocarbons with Mass No. 300 and 302 in Hard Coal Flue Gas. *Fresenius Z. Anal. Chem.* **1987**, *326*, 401-413.
(7) Simonsick, W. J.; Hite, R. A. Characterization of High MW Polycyclic Aromatic Hydrocarbons by Charge Exchange Chemical Ionization MS. *Anal. Chem.* **1986**, *58*, 2114-2121.
(8) Gerhardt, Ph.; Homann, K. Ions and Charged Soot Particles in Hydrocarbon Flames. *J. Phys. Chem.* **1990**, *94*, 5381-5391.
(9) Dias, J. R. *Handbook of Polycyclic Hydrocarbons, Part B*; Elsevier: New York, 1988.
(10) Schmidt, W. PAH Institut, Flustrasse 15, Greifenberg 8919, Germany.
(11) Dias, J. R. Constant-Isomer Benzenoid Series and Their Topological Characteristics. *Theor. Chim. Acta* **1990**, *77*, 143-162.
(12) Dias, J. R. Benzenoid Series Having a Constant Number of Isomers. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 251-256. *Ibid.* **1991**, *31*, 89-96.
(13) Dias, J. R. Constant-Isomer Benzenoid Series and Their Polyradical Subsets. *Theor. Chim. Acta* **1991**, *81*, 125-138.
(14) Sachweh, V.; Langhals, H. Synthesis of Pure Rubicene and Rubicene Derivatives. *Chem. Ber.* **1990**, *123*, 1981-1987.
(15) Wise, S. A.; Campbell, R.; West, W. R.; Lee, M. L.; Bartle, K. D. Characterization of Polycyclic Aromatic Hydrocarbon Minerals Curtisite, Idrialite, and Pendletonite Using HPLC, GC, MS, and NMR. *Chem. Geol.* **1986**, *54*, 339-357.
(16) Clar, E.; Robertson, J.; Schlogl, R.; Schmidt, W. PE Spectra of PNAs—Applications to Structural Elucidation of Circumanthracene. *J. Am. Chem. Soc.* **1981**, *103*, 1320-1238.
(17) Randić, M.; Trinajstić, N. Conjugation and Aromaticity of Corannulenes. *J. Am. Chem. Soc.* **1984**, *106*, 4428-4434.
(18) Gerson, F.; Knobel, J.; Metzger, A.; Murata, I.; Nakasuji, K. Radical Ions of Conjugated Polycyclic Hydrocarbons Containing Two Phenalenyl π-Systems. *Helv. Chim. Acta* **1984**, *67*, 934-938.
(19) Stein, S. E.; Fahr, A. High Temperature Stabilities of Hydrocarbons. *J. Phys. Chem.* **1985**, *89*, 3714-3725.
(20) Schaden, G. A Simple Synthesis of Pyracylene. *J. Org. Chem.* **1983**, *48*, 5385-5386.
(21) Slupek, S.; Kozinski, J. A. Determination of PAHs in Heavy Oil Flames by GC/MS. *Fuel* **1989**, *68*, 877-882.
(22) Wentrup, C.; Benedikt, J. Nitrile Imine and Carbene Rearrangements. *J. Org. Chem.* **1980**, *45*, 1407-1409.
(23) Tsunetsugu, J.; Tanaka, S.; Ebine, S.; Morinaga, K. Synthesis and Properties of Cyclohepta[*fg*]naphth[2,3-*a*]acenaphthylene-5,12-dione. *J. Chem. Soc., Perkin Trans. 1* **1988**, 1541-1545.
(24) Randić, M.; Trinajstić, N. On the Relative Stabilities of Conjugated Heterocycles Containing Divalent Sulfur. *Sulfur Rep.* **1986**, *6*, 379-430.
(25) Taylor, R. A Valence Bond Approach to Explaining Fullerene Stabilities. *Tetrahedron Lett.* **1991**, *32*, 3731-3734.
(26) (a) Scott, L. T. Thermal Rearrangement of Aromatic Compounds. *Acc. Chem. Res.* **1982**, *15*, 52-58. (b) Scott, L. T.; Roelofs, N. H. Benzenoid Ring Contraction in the Thermal Automerization of Acenaphthylene. *Tetrahedron Lett.* **1988**, *29*, 6857-6860.
(27) Balasubramanian, K. Enumeration of Isomers of Polysubstituted C_{60} and Application to NMR. *Chem. Phys. Lett.* **1991**, *182*, 257-262.
(28) Zahradnik, R.; Pancir, J. *HMO Energy Characteristics*; IFI/Plenum Press: New York, 1970.
(29) Dyker, G. Acenaphth[1,2-*a*]acenaphthylene: A Semi-benzenoid Hydrocarbon with Dienophilic Central Double Bond. *Tetrahedron Lett.* **1991**, *32*, 7241-7242.
(30) Dias, J. R. A Periodic Table for Polycyclic Aromatic Hydrocarbons. Isomer Enumeration of Fused Polycyclic Aromatic Hydrocarbons. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 15-22.
(31) Cyvin, B. N.; Brunvoll, J.; Cyvin, S. J. Notes on Fully Benzenoid Hydrocarbons and Their Constant-Isomer Series. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 72-78.

A Program for the Forward Generation of Synthetic Routes

JAMES B. HENDRICKSON* and CAMDEN A. PARKS

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

Received October 29, 1991

Following the skeletal dissection of a target structure, functionality and reactions may be generated either retrosynthetically back from the target to the necessary starting materials or in the forward direction from catalogue starting materials up to the target skeleton with various functionality, to be altered to the target itself after skeletal construction. The former represents the SYNGEN program; the latter is the FORWARD program described here. The FORWARD variation is important not only in being directed by available starting materials but also in solving the problem of deducing refunctionalization reactions as well as constructions.

The problem of automated synthesis design is basically one of combinatorics: the number of possible routes to any target is far greater than is generally appreciated. In principle it is simple for the computer to generate them all; the main task is to establish stringent criteria that allow for the selection of a rather small number of optimal routes.[1]

Our approach to automated synthesis design is centrally based on a criterion of *economy*. This dictates finding the shortest, most efficient routes to the target from real starting materials available commercially. Our developing theory of synthesis design logic takes the assembly of the target skeleton as the key to the selection of the best synthetic routes. This demands that we first separate the skeleton from its appended functionality and then derive the shortest, most efficient dissections of that skeleton to assemble it from a set of available starting material skeletons. The best syntheses are then the shortest ones, i.e., those that simply construct in the forward direction each of the dissected skeletal bonds without stopping to refunctionalize or repair functionality on the way. Such a sequence, of construction reactions only, has been regarded as an *ideal synthesis* and is the approach taken by the SYNGEN program.[1]

The first task of the SYNGEN program then is dissection of the skeleton to available starting material skeletons. This is followed by a search for the necessary functionality. This search for functionality concentrates on those starting materials that are most readily convertible to target, i.e., are closest to it in the synthesis tree. When starting materials are ideal, their conversion to target is easily found as a sequence of construction reactions, but when they need refunctionalization as well as joining en route, they will not be obvious in any retrosynthetic approach. Nevertheless, a catalogue of starting materials constitutes available information for the synthesis designer and so should be incorporated in any protocol. The

0095-2338/92/1632-0209$03.00/0

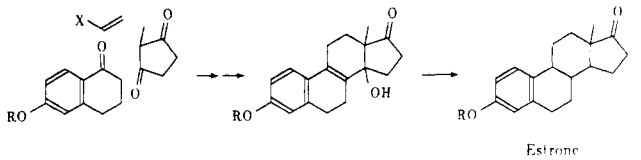


Figure 1. Torgov–Smith estrone synthesis.

Gelernter program SYNCHEM also has such a catalogue[2] but it is not actively used to direct the dissection of the target as it is in SYNGEN. This is also true of the LHASA approach[3] and that of Bersohn;[4] in these programs the catalogue is simply used, if at all, to recognize a starting compound when the retrosynthetic analysis comes upon it. Wipke reinvented the skeletal simplification of ref 1a with his SST program[5] to locate starting materials, but it is not clear whether it was ever incorporated into synthesis design.

In any case, in real syntheses there are in fact more refunctionalization reactions than constructions.[6] These may occur before, during, and/or after the sequence of constructions. These are reactions that repair the functional groups on available starting materials to ready them for the constructions, reactions that alter functionality from the product of one construction to prepare it for the next, and reactions at the end that refunctionalize the final constructed skeleton to make the desired target.

The addition of prior refunctionalization of starting materials to the SYNGEN program was a relatively easy task,[7] but we were also interested in finding a way to incorporate the other modes of refunctionalization, particularly those that occur at the end of the synthesis. Many syntheses use a last step or two to remove *dummy* functional groups, i.e., those that are included to facilitate constructions en route but are not present on the target itself. This usually occurs when the target exhibits no functionality around the constructed skeletal bonds. Such routes cannot be deduced backwards from the target since these dummy groups leave no trace of their presence in the target structure itself. A case in point is the elegant synthesis of estrone summarized in Figure 1;[8] the synthesis is an ideal one to the intermediate, which is then refunctionalized to estrone. However, retrosynthetic deduction of this route from the structure of estrone itself is very unlikely with present programs.[1-4] Refunctionalization at the end of the skeletal construction sequence is also important in syntheses aimed at not just one target but a family of targets of the same skeleton with various attached functionality.

These considerations led us to develop the FORWARD program, in which one allows a set of all starting materials of the correct skeletons to proceed through a sequence of skeletal bond constructions to obtain the target skeleton bearing any functional groups that may naturally arise from the construction sequence. If these groups can at the end be refunctionalized easily to those of the target, such sequences would be accepted as synthesis pathways. In both SYNGEN and FORWARD the first step is to dissect the skeleton to establish the best pieces (starting skeletons) from which to assemble the target skeleton, and also the order in which to link them together.

The fully convergent assembly is the most efficient way to put together the starting material skeletons;[9] this is readily found by cutting the target skeleton into two parts and then cutting these two intermediate skeletal fragments into two more fragments each. Only one or two bonds may be cut each time, assuring that no more than one ring is cut when dividing a skeleton in two. This generates *bondsets*, i.e., sets of skeletal bonds cut which must be constructed in the synthesis: first-level bondsets from the target to the two intermediates, and second-level bondsets to the two pairs of skeletons which join to form those intermediates.

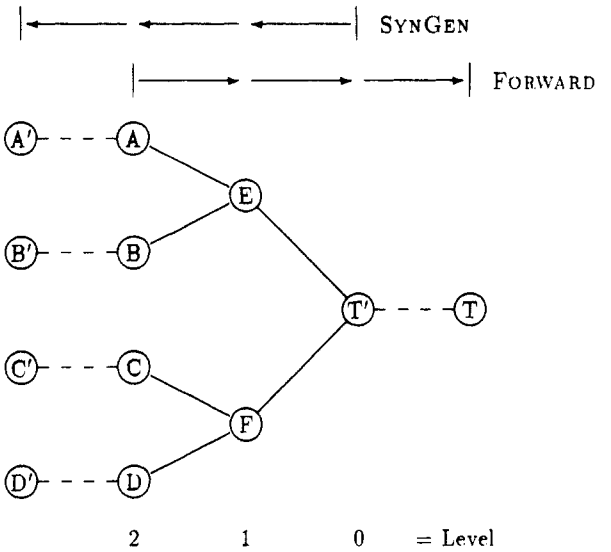


Figure 2. Comparison of retro and forward route generation. Solid lines indicate constructions; dashed lines indicate possible refunctionalizations.

Each time a new skeleton is generated in these dissections it is looked up in the catalogue of available starting material skeletons. Bondsets are flagged for priority at first level if an intermediate skeleton is found to be available, but are deleted at second level if all four resultant skeletons are *not* found in the catalogue. This protocol is stopped at the second level of dissection, where the four found skeletons are taken to be starting material skeletons. This assumes that enough routes can be found from these four starting skeletons, and these routes must of course be the shortest ones; further dissection only makes longer routes. The basis of the assumption is that a C_{20} target would have four starting materials averaging five carbons, and the experience with our starting material catalogue shows that the functional variety on nonaromatic starting materials falls off rapidly above about five carbons. Larger targets frequently have aromatic rings, for which larger starting materials are also available, so that the four starting material cutoff remains reasonable.

The point of this initial skeletal dissection is to locate all convergent assemblies of the target, from real starting skeletons, requiring construction of no more than six skeletal bonds (three cuts with at most two bonds per cut). The next step is to generate the appropriate functionality on the skeletons to minimize the actual number of chemical steps in any route.[10]

While the central body of the route consists only of sequential constructions, without refunctionalizations, we may consider refunctionalization steps either at the beginning or at the end of the construction sequence. To do the former we start with the target functionality and proceed retrosynthetically back through each sequential bond of a bondset, generating the required construction reactions and the functionality necessary to create each one. This will ultimately arrive backwards at the four starting skeletons and derive the functionality on them required to initiate the series of construction steps. If the necessary functional groups are not found in the catalogue for those starting materials, they may be refunctionalized from actual starting materials in the catalogue with the same skeleton. This is the retrosynthetic approach on which the program SYNGEN has been built.[1,7]

The alternative approach is to gather, for each bondset, all the starting materials in the catalogue which have the required starting skeletons and join them in the forward, or synthetic, direction toward the target skeleton. This is the basis of the FORWARD program. It implies that the pairs of starting materials will be used to generate sequential constructions forward, through defined intermediates, to products which have

the target skeleton but may have different functional groups, and so may then be refunctionalized to the target itself at the end of the synthesis.

The two approaches are illustrated in Figure 2, a generalization of the synthesis tree, with circles for compounds, lines for conversions, and the synthetic direction from starting materials at the left (A–D) to target (T′ or T) at the right. SYNGEN proceeds retrosynthetically from a given target T′ to generate two intermediates (E and F), from which it then generates the four starting materials (A–D). If these are not found in the catalogue, they may be created by prior refunctionlization from real starting materials of the same skeleton (A′–D′). In the FORWARD program, pairs of starting materials (A–D) from the catalogue are joined all ways into intermediates (E and F), and these are then joined to form target dervatives (T′), which are within a given number of refunctionalization steps from the target (T). The two programs will find the same set of *shortest* routes to a given target, those routes being the ones which involve no refunctionalizations of starting materials or products. The programs differ in the routes they generate that require refunctionalization: SYNGEN refunctionalizing *before* the construction sequence to fix the starting materials, FORWARD refunctionalizing *after* to fix the products.

The FORWARD program uses the same definitions of compounds and reactions as in SYNGEN, identifying the reactive strands out from each end of each bond constructed and generating all possible half-reactions at each end. The chemistry generator determines which half-reactions are to be retained as valid by applying lists of activating and restricting chemical conditions. It then matches pairs of valid half-reactions, nucleophile to electrophile.[7] In SYNGEN this is done retrosynthetically, but in FORWARD the half-reactions are generated in the forward direction, from reactants to products.

PROCEDURE AND COMBINATORICS CONTAINMENT

As outlined above both the SYNGEN and FORWARD programs begin by stripping the target down to its skeleton and dissecting it twice (in two levels as in Figure 2) to create four starting skeletons (A–D) available in the catalogue. This creates a set of convergent bondsets. In the FORWARD program all representatives of each starting skeleton are now collected, and these are joined pairwise by generating all possible construction reactions which can be initiated by their functional groups to create the designated skeletal bonds. This generates a set of second-level reactions to the intermediate compounds, i.e., A + B → E and C + D → F. These intermediates are similarly joined in all possible ways, generating a set of first-level reactions, to form products with the target skeleton, each bearing the natural functionality resulting from such constructions.

Each catalogue compound with a starting material skeleton must be mapped onto the target skeleton all possible ways. For example, an asymmetrically functionalized cyclohexane may be linked by one bond at any of its six carbons, so it will have six distinct mappings. Therefore, the number of starting materials to be used is much increased by these permutations of compounds with symmetrical skeletons.

In principle this pairwise combination of all permutations of all starting materials leads to an unwieldy combinatoric explosion, yielding very large numbers of variously functionalized target skeletons, most very far removed from the desired target itself. In order to contain this combinatoric explosion, we must apply a measure of whether the reactions generated lead in the direction of the functionality of the desired target. Such a measure is available in the *reaction distance*,[11] in effect the number of unit reactions or reaction steps, which can be calculated between any two compounds with their skeletal atoms mapped. This calculated reaction distance is essentially the same as the *chemical distance* between any two structures as derived from their adjacency matrixes in the Ugi method;[12] the two distance calculations have been compared in ref 13. This reaction distance allows us to see the degree of functional similarity between any mapped pair of compounds. It therefore allows the program to delete at the outset those starting material permutations which are functionally too distant from the target. The same calculation on the intermediates will allow for deletion of those which are likewise too distant from the target.

The reaction distance Δ between two compounds is calculated as

$$\Delta = \sum_i (|\Delta h_i| + |\Delta z_i|)$$

in which Δh_i and Δz_i are the differences in the numbers of hydrogens and heteroatom functional attachments, respectively, between the two structures at carbon i. The number of unit reactions (steps) for one structure to pass to the other is then $\Delta/2$.

Each starting material must first be mapped to its carbons in the target skeleton in all nonredundant permutations, and the reaction distance (Δ) to the target calculated for each permutation. Those carbons which bear bondset bonds (those which are to be constructed in assembling the target skeleton) require an average of $\Delta = 2$ each; if there are L links or bonds to be made to a starting material, the reaction distance to target will be $2L + x$, in which $2L$ represents the obligatory construction steps (skeletal bonds to make) and x represents the extra steps, i.e., refunctionalizations, allowed in passing this starting material to the target. Values of x can be set by the user, both for the starting materials and for the intermediates.

Furthermore, the user may designate whether the program is to accept all reactions, only nondivergent reactions, or only convergent reactions en route to the target: nondivergent reactions show no increase in Δ (to target) in the reaction, while convergent reactions show a decrease in Δ. To focus the program further the user may also set limits on how far the generated products (T′ in Figure 2) may be from the target itself in terms of refunctionalization steps, i.e., by setting a maximum Δ for T′; that product for which $\Delta = 0$ is of course the target compound itself.

Once all the mapped permutations of all starting materials of correct skeleton have been generated, only those with Δ within the user-supplied limit are retained. The program then combines these starting materials to form intermediates. This is done as it is in SYNGEN but in the forward instead of retrosynthetic direction, using half-reactions on each end of a bond being constructed and pairing them as nucleophile–electrophile across the bond.[7] Those constructions for which the change in Δ for the reaction fits the user-supplied criterion (i.e., 'keep only convergent reactions') generate intermediates (E and F); only those intermediates which are within the user-supplied limit ($2L + x$) are retained. The intermediates are combined in the same manner to form products, which must also be within supplied Δ limits. This will result in a set of derivatives (T′) of the given target bearing the functionality which naturally results from the successive constructions, and normally only two steps or less from the given target ($\Delta \leq 4$).

RESULTS AND DISCUSSION

In using FORWARD the user first draws the target with the same fast, fluent drawing procedure used in SYNGEN. The

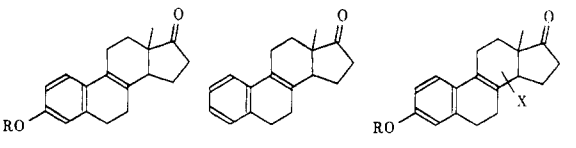
user is then asked to define the limits allowed for the generation. The five user-definable limits are the maximum amount of "excess" functionality (x in $2L + x$) for starting materials; the maximum "excess" functionality for intermediates; the maximum Δ for products; the allowable change in Δ over the course of a reaction (retain only convergent, nondivergent, or all reactions); and the minimum size for a first-level skeleton. The last of these (a limit also used in SYNGEN) is a lower limit on the size of the smaller piece after the first skeletal cut of the target skeleton. The default size for the smaller of the two first-level skeletons is taken as one-fourth of the skeletal atoms in the target.[14] Because of the implicit combinatorics, the required time for the program (written in Fortran and run on a Micro VAX 3500) is of course very dependent on the limits set. Estrone required 2 h of CPU time with the default limits of $\Delta_{max} = 2L + 2$ for both starting materials and intermediates, and $\Delta_{max} = 4$ for target products, convergent reactions, and all first-level skeletons at least one-fourth the size of the target skeleton.

The output is displayed so that the user may choose to see collected statistics on the numbers of bondsets, starting materials, intermediates, and products which are generated, broken down by their Δ values. He may also choose to see the bondsets or any of the compound structures as well as the reactions generated. There is a powerful set of retain and delete options for use on bondsets, compounds, and equal Δ families of compounds; these options can be used to help the user prune the output to manageable size.

Table I shows for each target a number of program runs with the values of the five user-defined limits, and the results in the form of numbers of starting materials, intermediates, products, bondsets, and reactions. The numbers in the Results section of the table are indicative only of how many compounds, bondsets, or reactions appear in syntheses which were kept. For example, in line 1 of Table I we see that the output set for estrone with all default limits has 38 second-level bondsets; however, the program started with 78 second-level bondsets which had starting materials of $\Delta \leq 2L + 2$. Similarly, there were likely many more than 320 products generated, but only those of $\Delta \leq 4$ were kept.

When estrone is examined with default values for the limits (entry 1 in Table I), the immediately striking feature is the very large number of starting material permutations which could be incorporated (807) and a correspondingly large number of intermediates (878). Although only six first-level bondsets were employed to join the pairs of intermediates (E and F), there were 5768 reactions found to do it, leading to 320 variants of estrone with the same skeleton and differing from estrone by no more than two refunctionalization steps ($\Delta = 4$). Among these is estrone itself ($\Delta = 0$) and also the intermediate of the Torgov–Smith estrone synthesis (Figure 1) ($\Delta = 4$). Of the 320 products, 24 have a Δ of 2 and 295 have a Δ of 4. This dramatizes how many combinations are possible for varying estrone this way, i.e., with two double bonds variously placed, one double bond and one heteroatom attachment, or two heteroatoms (including ketones). This in turn points out how nearly impossible it would be to *find* these synthetic routes, which remove dummy functional groups, by deducing backwards directly from the estrone structure, in the manner of SYNGEN and other retrosynthetic protocols.[1-4]

Exploring the data set for entry 1 of Table I we find that when we examine only those routes which lead to estrone itself (the $\Delta = 0$ product) there are still 141 starting material permutations and 17 intermediates joined in 18 first-level annelations using four of the six first-level bondsets. This dramatic reduction from the default setting is all that remains for the ideal syntheses, without refunctionalizations. When we make all the Δ limits as tight as possible (entry 2: no excess



1 2 3–5

Figure 3. Selected FORWARD-generated estrone derivatives.

functionality for starting materials and intermediates, keep only the target compound, and retain only convergent reactions), we find that we get the same set of results as when we looked at the routes leading only to estrone itself in the data set for entry 1. If instead, using the output from the default case of entry 1, we only restrict the *bondsets* at each level (entry 3) to just those of Figure 1, the variety falls to just 28 reactions total but still creates five estrone derivatives from 13 starting materials on the skeletons of Figure 1. The five derivatives are shown in Figure 3.[15] Derivative **1** has only the extra double bond; derivative **2** also lacks the aromatic hydroxyl of estrone. The intermediate shown in Figure 1 is the same as one of the derivatives corresponding to structure **3** of Figure 3, and one of the reactions forming that derivative is in fact that of the Torgov–Smith synthesis.[8]

We may further examine the effects of the five definable limits on the combinatorics for the chemistry of estrone synthesis. We find first (entry 4) that allowing the first dissection cut to remove one skeletal piece of as little as one carbon has, as expected,[14] very little effect on the final numbers. Opening the possibilities for starting materials to $x = 90$ has very little effect (entry 5); this is due to the fact that starting materials with relatively high Δ are unlikely to be usable for the formation of intermediates with $\Delta \leq 2L + 2$. Similarly, opening just the limit on intermediate Δ also has very little effect on the results relative to the default case; this is due to the difficulty of forming high Δ intermediates from low Δ starting materials, especially when only convergent reactions are allowed. Opening the limits on the product Δ causes the program to fail because the number of reactions generated exceeded the 60 000 maximum set for the program.

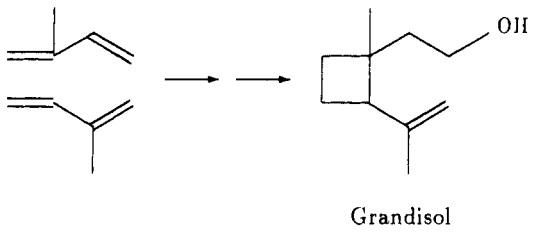
Narrowing the estrone limits to $x = 0$ for starting materials and intermediates and $\Delta = 4$ for products (entry 6) does not reduce the numbers much, but the same values are obtained for the reactions to estrone itself, and again the suppression to just the bondset of Figure 1 leaves two of the five estrone derivatives shown in Figure 3 (**1** and the Torgov–Smith intermediate). In this case the effect of also allowing *all* reactions, i.e., accepting any change in Δ on reaction, was explored (entry 7); the numbers were found to increase but not by much. These results in Table I give an impression of the degree of interrelation among the five user-definable limits.

Table I shows results from other targets to demonstrate further the utility of the Δ limits in containing the combinatoric explosion. Since estrone was too big a target to follow the various possible ways of opening up restrictions, the smaller 10-carbon monoterpene grandisol was chosen as target. Here we see (entry 8) the default limits lead to 179 products in 4960 first-level reactions; when we look at the routes leading only to the target compound we find 29 starting materials, 10 intermediates, 42 second-level reactions, and 15 first-level reactions. Comparing these numbers to those of entry 9 we find that some of those routes required some excess functionality in either the starting materials or the intermediates. Contrast this to the case of estrone in which, as we saw above, none of the routes to the target generated by the use of the default values for the Δ limits involved starting materials or intermediates with excess functionality. Again the relatively small output for the ideal synthesis is apparent. The effect of removing the minimum size limit on first-level skeletons is again very small, as seen in entry 10.

Table I. Representative Output from the FORWARD Program

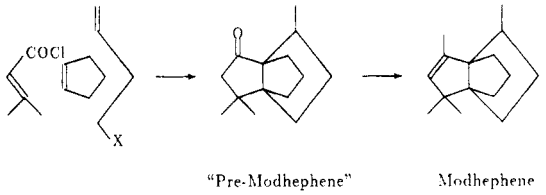
		restriction settings					results: numbers of						
		values of x[a]		Δ[a]		level 1							
no.	target	SMs	ints	prods	rxns	frag[b]	SMs[a]	ints[a]	prods[a]	BS-2[c]	BS-1[c]	rxns-2[c]	rxns-1[c]
1	estrone	2	2	4	con	def	807	878	320	38	6	8462	5768
2	estrone	0	0	0	con	def	141	17	1	25	4	332	18
3[d]	estrone	2	2	4	con	def	13	6	5	1	1	14	7
4	estrone	2	2	4	con	≥1	813	885	320	40	8	8477	5815
5	estrone	90	2	4	con	def	827	884	321	38	6	8536	5788
6	estrone	0	0	4	con	def	488	310	293	36	5	3703	4027
7	estrone	0	0	4	all	def	501	332	293	36	5	4014	5055
8	grandisol	2	2	4	con	def	180	376	179	9	4	1110	4960
9	grandisol	0	0	0	con	def	25	9	1	8	4	38	14
10	grandisol	2	2	4	con	≥1	193	498	179	16	6	1475	6583
11	grandisol	90	2	4	con	def	200	451	190	9	4	1253	5277
12	grandisol	2	90	4	con	def	186	437	179	9	4	1620	5072
13	grandisol	2	2	90	con	def	200	474	3316	9	4	1345	42 509
14	grandisol	0	90	90	con	def	114	263	581	9	4	1018	6468
15	grandisol	0	90	90	nondiv	def	120	490	2216	9	4	1646	21 746
16	grandisol	0	90	90	all	def	120	509	2278	9	4	1687	22 788
17	modhephene	2	2	4	con	def	171	313	228	8	3	824	2809
18	modhephene	0	0	0	con	def	24	10	1	6	3	25	18
19	modhephene	2	2	4	con	≥1	189	325	231	11	6	836	3114
20	modhephene	2	2	4	nondiv	def	174	334	230	8	3	907	3199
21	modhephene	2	2	4	all	def	174	334	230	8	3	908	3199
22	"pre-mod"	2	2	4	con	def	156	248	142	9	3	722	1702
23[d]	"pre-mod"	2	2	4	con	def	46	17	56	1	1	42	193
24	reserpine	2	2	4	con	def	129	135	291	14	2	370	407
25[d]	reserpine	2	2	4	con	def	41	17	15	6	1	48	19

[a] SM = starting materials; int = intermediates; prod = target-skeleton products (T′). [b] Default minimum size of first-cut skeletons is one-fourth target size. [c] Bondsets at 1st and 2nd level; reactions at 1st and 2nd level. [d] Original defaults shown but further pruned as described in text.



Grandisol

Figure 4. Grandisol synthesis.



"Pre-Modhephene" Modhephene

Figure 5. Modhephene synthesis.

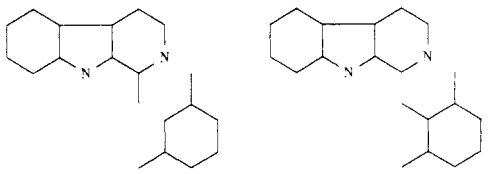


Figure 6. Reserpine bondsets.

The next three entries show the effects of removing the Δ limits on each of the three types of compound: as was the case with estrone we see that removing the Δ limit on products has the greatest effect on the size of the results set, generating over 42 000 first-level reactions leading to over 3000 products with Δ values of up to 10. The last three entries for grandisol (14–16) explore the effect of relaxing the restrictions on the allowable change in Δ over the course of a reaction: passing from converging to nondiverging reactions causes a large increase, but since the starting material restriction has remained tight, the possibilities for more reactions which actually diverge from target direction are quite limited.

For comparison purposes, the shortest literature synthesis of grandisol[16] is just two steps, a photochemical dimerization which goes in low yield but is readily refunctionalized in one further step to the target, as shown in Figure 4. When the output set for the default limits is reduced to products of $\Delta \leq 2$ and the same bondset is selected, several close analogues of this synthesis are found.[17]

The several entries (17–21) for the sesquiterpene modhephene are included in Table I to show that the same general trends are maintained and are not largely a function of the nature of the target. The several syntheses of modhephene in the literature add one or more of the methyls at the end, and so retrosynthetically do not fit the convergent dissection described above. Two of them, however, do create the basic skeleton in an "ideal" fashion essentially as shown in Figure 5.[18] When this "pre-modhephene" is entered in the FORWARD program the results are shown in Table I as entries 22 and 23, and the ideal syntheses include the one summarized in Figure 5, easily found by selecting only that bondset.

Finally, the output for reserpine is summarized in Table I (entries 24 and 25) to show that the program also handles skeletons containing heteroatoms and that the results are consistent. Since the target now has 22 skeletal atoms, the possibilities for a convergent dissection in just two levels are much diminished, there being only two bondsets at first level which are successful. These are shown in Figure 6; it may be noted that the common bondset for reserpine which cuts three skeletal bonds to isolate tryptamine (β-indolylethylamine) as a starting material is not allowed here because of the stricture of cutting no more than two bonds to isolate two intermediates (E and F). When we look at only those products which contain actual indole or dihydroindole structures we find the results shown in entry 25, the single first-level bondset of which is that on the left in Figure 6.

CONCLUSIONS

This preliminary work with the FORWARD program shows that the idea of generating synthetic routes in the forward or synthetic direction from convergent bondsets, by combining all possible starting materials, is workable as long as the reaction distance is used to focus the generation on paths which arrive at the target with no more than a few steps of final refunctionalization. With the commands available in the output routine for deleting unwanted families of routes, the often large numbers of reactions generated may be rapidly and easily pruned down to focus further on just a few optimal syntheses. Figure 2 indicates that the *shortest* syntheses will be those with no refunctionalizations at either end of the ideal construction sequence and so should be found by both SYNGEN and FORWARD; in the several targets illustrated here this was found to be the case, cf., entry 2. Now that the procedure has proven to be effective, we will rewrite the program to run much more efficiently and rapidly.

The logical basis for synthesis design which we are implementing consists of a prior skeletal dissection of the target to the best convergent assemblies of that target from available starting material skeletons. These skeletal assemblies from ordered bondsets are then elaborated with functional groups to afford a sequence of construction reactions, generated retrosynthetically by SYNGEN and synthetically by FORWARD. Extending the possibilities by incorporating refunctionalizations is done by SYNGEN on the starting materials (A–D in Figure 2) before the central construction sequence, and by FORWARD on target skeleton derivatives (T′) after the central sequence. Our plan now is to extend the procedure in the same way to refunctionalize *during* the central construction sequence, on intermediates E and F in Figure 2. Ultimately, a more flexible synthesis route generation will result by incorporating all three modes of refunctionalizations in a single synthesis generator.

ACKNOWLEDGMENT

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

REFERENCES AND NOTES

(1) (a) Hendrickson, J. B.; Braun-Keller, E.; Toczko, A. G. A Logic For Synthesis Design. *Tetrahedron* **1981**, *37*, 359–370. (b) Hendrickson, J. B. Approaching the Logic of Synthesis Design. *Acc. Chem. Res.* **1986**, *19*, 274–281.

(2) Gelernter, H.; Bhagwat, S. S.; Larsen, D. L.; Miller, G. A. Knowledge-Base Enhancement via Training Sequence: The Education of SYNCHEM. In *Computer Applications in Chemistry*; Heller, S. R., Potenzone, R., Eds.; Elsevier: New York, 1983; pp 35–59.

(3) Corey, E. J.; Long, A. K.; Rubenstein, R. Computer-Assisted Analysis in Organic Synthesis. *Science* **1985**, *228*, 408–418.

(4) (a) Bersohn, M.; Esack, A.; Luchini, J. A. Computer Representation of Synthetic Organic Reactions. *Comput. Chem.* **1976**, *2*, 103–107. (b) Bersohn, M.; Esack, A. Computers and Organic Synthesis. *Chem. Rev.* **1974**, *76*, 269–282.

(5) Wipke, W. T.; Rogers, D. Artificial Intelligence in Organic Synthesis. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 71–81.

(6) On average in published syntheses there are usually twice as many, leading one to suppose that these may not be the most efficient paths possible (cf. ref 1).

(7) Hendrickson, J. B.; Toczko, A. G. SYNGEN Program for Synthesis Design: Basic Computing Techniques. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 137–145.

(8) (a) Ananchenko, S. N.; Torgov, I. V. New Syntheses of Estrone, *d*,1-8-iso-estrone and *d*,1-19-nortestosterone. *Tetrahedron Lett.* **1963**, 1553–1558. (b) Smith, H.; et al. Totally Synthetic (±)-13-Alkyl-3-hydroxy and Methoxy-gona-1,3,5(10)-trien-17-ones and Related Compounds. *Experientia* **1963**, *19*, 394–396. (c) Smith, H.; et al. Totally

Synthetic Steroid Hormones. Part 1. Estrone and Related Estrapolyenes. J. Chem. Soc. 1963, 5072-5094.
(9) Hendrickson, J. B. Systematic Synthesis Design. 6. Yield Analysis and Convergency. J. Am. Chem. Soc. 1977, 99, 5439-5450.
(10) Each full bondset, at both levels, defines an ordered sequence of constructions, hence a family of routes differing only in the involved functional groups. The simplest overall description of any synthetic route is just its ordered bondset, i.e., simply the sequence of skeletal bonds constructed in the course of the synthesis.
(11) Hendrickson, J. B.; Braun-Keller, E. Systematic Synthesis Design. 8. Generation of Reaction Sequences. J. Comput. Chem. 1980, 1, 323-333.
(12) Wochner, M.; Brandt, J.; von Scholley, A.; Ugi, I. Chemical Similarity, Chemical Distance, and Its Exact Determination. Chimia 1988, 42, 217-225.
(13) Hendrickson, J. B. Descriptions of Reactions: Their Logic and Applications. Recl. Trav. Chim. Pays-bas 1992, in press.
(14) Allowing the first dissection of the target to cut off one- and two-carbon pieces generates many first-level bondsets but leaves the other intermediate piece so big that it is very unlikely that any further second-level dissection will then find real starting materials, and so such cuts are almost always wasteful.
(15) Estrone itself does not appear here since it is only successfully made by other bondsets.
(16) Billups, W. E.; Cross; Smith, C. V. A Synthesis of (±)-Grandisol. J. Am. Chem. Soc. 1973, 95, 3438-3439.
(17) For reasons involving the restrictions on generating these reactions[7] the exact dimerization does not appear.
(18) (a) Oppolzer, W.; Battig, K. A. Short and Efficient Synthesis of (±)-Modhephene by a Stereoelectrically-Controlled Ene-Reaction. Helv. Chim. Acta 1981, 64, 2489-2491. (b) Schostarez, H.; Paquette, L. A. Highly Stereocontrolled Synthesis of [3.3.3]Propellane Sesquiterpenes. (±)-Modhephene and (±)-Epimodhephene. Tetrahedron 1981, 37, 4431-4435.

Algorithm for Selecting the Parent Structural Unit of a Ring–Chain Assembly

SCOTT DAVIDSON*

Computer Data Systems, Inc., One Curie Court, Rockville, Maryland 20850

Received November 18, 1991

Selection of the parent structural unit in naming ring–chain assemblies by computer methods can be difficult. The IUPAC rules are sufficiently general to permit individual judgments in some cases. Nodal nomenclature employs a quite specific but very lengthy set of rules to give unique names. A recently reported improved method for uniquely identifying alkanes defines the main chain as the chain with the least complex side chains. This definition readily extends to ring–chain assemblies when rings are included, applies regardless of parent-unit size, and is consistent with the IUPAC guidelines and examples given, such as the preference for diphenylmethane over benzyl benzene in naming Ph–CH_2–Ph. An iterative procedure for selecting the parent unit and a simple method for linking units are described as part of a general skeleton-naming computer program.

INTRODUCTION

In organic chemical nomenclature, the selection of a parent structural unit is essential to naming an assembly of rings and chains. The selection process is complicated by the fact that in an assembly the individual units are independent with respect to simplification. In a single-unit carbon skeleton, selection of a main chain or ring usually removes most carbon atoms from the naming process, whereas with an assembly complete identification of one unit provides little information about the others, which can have arbitrary sizes and shapes. IUPAC nomenclature (Rule A-61.2)[1] provides two flexible, highly intuitive general guidelines for selecting the parent unit: "(a) the maximum number of substitutions into a single unit of structure; (b) treatment of a smaller unit of structure as a substituent into a larger". However, as reported by Goebels et al.[2] in reference to operation of the AUTONOM structure-naming program, such guidelines are not very well suited to computer implementation, but on the other hand the fixed rule of rings senior to chains employed by CAS can lead to overcomplicated names for some simple compounds.

Nodal nomenclature[3] provides a specific sequence of rules for selecting parent units (modules) with size as the highest level qualifier, followed by a set of 16 rules for numbering the remaining units. The basic concept of defining the entire skeleton first as single-bonded carbon and then overlaying bond and atom types is well suited to computerization. The use of the perimeter ring (largest ring) as the basic structural unit of multibridged systems in place of the bicyclic skeleton of the extended von Baeyer method simplifies structures and also avoids the proliferation of numbering schemes that became standard for common ring systems, bridged rings, and spiro structures before systematic nomenclature was developed.

Nomenclature rules commonly use linear sequences of tiebreaker tests to arrive at a unique name. The danger in this approach is that any one test that relies on incomplete information can cause the entire series to fail. For example, selecting the longest chain in an alkane requires examination of the complete structure. However, the first step in the IUPAC tiebreaker series (Rule 2.6a) for equal-length candidate main chains is to select the chain with the most side chains. This is a simple test—especially for a computer—because it relies on blind groping along a path. However, it ignores side-chain detail, allowing, for example, two methyl groups on one chain to override any single but highly complex side chain on the other. The first paper[4] shows how this can lead to nonunique names and describes a recursive algorithm for selecting an alkane main chain by side-chain complexity minimization. This alternative approach encodes and sorts structural details from the bottom up by depth-first search to give an ordered list (simple side chains + smaller lists encoding complex units). This can be compared with other lists representing alternate configurations from the top down as a tiebreaker series by symbolic substitution. The first value compared is the overall size of the skeleton, which of course is the same for all configurations, but permits a database of skeletons to be ordered in the same manner beginning with size. Recording details and then breaking ties at the lowest levels first ensure that all information is utilized in making decisions at the next higher level, etc. The main chain is

* Address all correspondence to this author at 240 Manor Circle 2, Takoma Park, MD 20912.

0095-2338/92/1632-0215$03.00/0