Fast Generation of an Alkane-Series Dictionary Ordered by Side-Chain Complexity

Scott Davidson†

240 Manor Circle #2, Takoma Park, Maryland 20912

Received September 5, 2001

Selecting the main chain of an alkane as the path that yields the least complex side chains without the maximum-length constraint leads to an efficient generating algorithm representable as nested binary trees. The largest side chain required to specify an N-carbon alkane becomes (N-1)/3. This allows 3.8 million C1−C22 alkanes to be coded for name translation in dictionary order, using an alphabet of 33 C1−C6 alkyl groups also ranked by complexity. The generating process produces reversible isomer codes already in canonical order, making the computation rate in isomers per second inverse linear with N and much faster than reported rates for other structure generators.

BACKGROUND

This paper continues previous studies[1−3] into applications of side-chain complexity minimization algorithms for naming and coding carbon skeletons of alkanes and ring-chain assemblies. Here the focus is on fast canonical generation of the alkane series. The original algorithm utilized for naming alkanes (ref 1) solved problems associated with equal-length main chain candidates (main chain ties). Specifically, there is a class of alkanes—beginning at C20—that passes the sequence of four IUPAC tiebreaker tests without yielding a unique name. These tests rely on side-chain count and locant rules that ignore structural differences. A simplifying modification presented here avoids these problems altogether and leads to an efficient method for generation and naming of alkanes as nested binary trees. This is NOT the common use of chemical trees for connecting atoms but rather for logically connecting all structural alkane isomers of a given size in a manner that minimally increments the alkyl-group complexity of the current isomer to obtain the next.

Side chain complexity as used here is derived from Rule 2.3 of the original IUPAC Rules[4] (1957): "If two or more side chains of different nature are present, they may be cited (a) in order of increasing complexity or (b) in alphabetical order." The complexity option was abandoned in subsequent revisions.[5] Former rule 2.3a(iii) says—as part of a tiebreaker series comparing side chains of equal size/length alkyl groups—"The less complex is the one whose longest substituent has the lower locant." This can be generalized by replacing "longest" with "higher ranking" to incorporate size and branching differences. Since a side chain is rooted at the main chain, this rule in essence says that lower complexity means more treelike, i.e. the largest branches are closest to the trunk and the ground. The complexity citation rules were completely separate from the main chain selection rules but were used in ref 1 as the basis of a recursive algorithm that ordered fragments by the ranking sequence (size, length, locant) in a depth-first search. Accounting for all atoms before selecting the main chain always produces unique names.

† Corresponding author phone: (202)693-2932; e-mail: davidson@uis.doleta.gov.

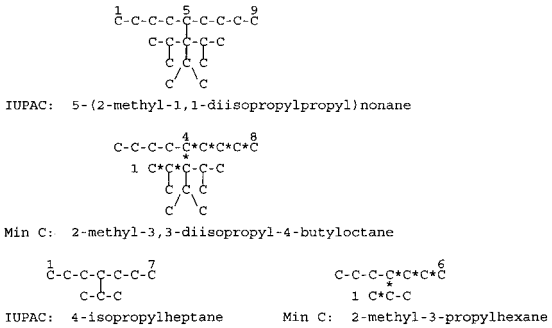

Figure 1. Alkanes named by unconstrained alkyl complexity minimization.

The concept of the central carbon atom played a key role in the first successful enumeration of alkane isomers by Henze and Blair.[6] Most recently, Bytautas and Klein[7] have cited the close relationship of this concept to substituted methane nomenclature. During the mid 1950s this nomenclature was taught as an acceptable alternative to the newly introduced IUPAC Rules in introductory organic chemistry.[8] It is still used occasionally today, a well-known example being diphenylmethane (Ph-CH_2-Ph). Note that in this case the parent unit is not the largest. However, the alternative name benzylbenzene creates a side chain that is larger than the parent unit benzene. The substituted methane name minimizes side chain complexity. In what follows, a similar idea is applied to alkanes.

The familiar longest chain rule tends to reduce side chain complexity by minimizing the complement of alkane size and length—the number of side-chain carbon atoms. However, this rule ignores the distribution of the carbon atoms, eventually allowing all of them to pile up in a way that creates side chains that are larger than the main chain and unnecessarily complex. Consider the C19 example shown in Figure 1: Note that the 10-carbon side chain (under the IUPAC Rules) is larger than the main chain. However, by routing the main chain so as to minimize side-chain complexity without regard to length, this big side chain is broken up into smaller segments. The largest side chain is reduced from 10 to 4 carbons and its locant from 5 to 4, but only 1 unit is lost from the main chain length (path marked with '*'). In the second simpler, more typical example the

10.1021/ci010094b CCC: $22.00
Published on Web 03/08/2002

Table 1. Maximum Size and Number of Alkyl Groups Required To Specify All Alkanes up to a Given Size without the Longest Chain Constraint

#C<=	total isomers	-# max alkyls =	# w alkyl size	<=	=	#alkyls∧p
10	150	4 C3	146	2	2	7.19
13	1466	20 C4	1446	3	4	5.25
16	18030	120 C5	17910	4	8	4.72
19	251731	969 C6	250762	5	16	4.48
22	3807434	10660 C7	3796774	6	33	4.33

first non-IUPAC main chain occurs in a 10-carbon alkane, where the more complex isopropyl group is reduced to a propyl group on a lower locant. In both cases complexity is also reduced by moving the largest side chain closer to the base of the main-chain tree trunk.

In ref 1, alkanes and their side chains are compared recursively in the series (size, length, locant). Here the elimination of the length constraint from main chain selection also eliminates the main-chain tie problem but, more importantly, leads to a dramatic reduction in the number of side chains that must be considered in generating and naming alkanes. For alkanes, the largest side chain (s) required to name structural isomers increases by one atom at each 3s+1 size (e.g. C10, C13 etc.), forming trialkylmethane isomers that have equal-sized alkyl groups. Routing the main chain through any two groups always leaves the third as a side chain. The number of isomers of this type is given by the formula for the number of combinations of n distinct objects taken r at a time with repetition: C(n+r-1,r). Table 1 shows the maximum number of isomers that can be specified by a given number of side chains. For example, there are 1466 C1−C13 isomers. Subtracting the 20 combinations of the 4 C4 (butyl) isomers of $(C4)_3C$ gives 1446 isomers that can be specified by just four side chains: methyl, ethyl, propyl, and isopropyl. Also shown is the "power" of 4 (side chains) as the exponent giving 1446. This value exceeds 4.0 for alkanes into the billions of isomers. The table also shows that 3.8 million C1−C22 alkanes can be specified with 0−33 C1−C6 side chains. All of these and 36 of 39 C7 side chains have only methyl and ethyl branches.

ALKANE GENERATION

Most of the literature on isomer generation has focused on the enumeration of specific structural classes. Very little attention has been given to the sequence of generation and its possible relationship to nomenclature. A notable exception is the 1981 paper of Knop et al.[9] that introduced the N_tuple code in this journal. Here the 159 undecane isomers are displayed in the order generated. Contreras has extended this code to generate cyclic and other increasingly complex structures in a series of recent papers.[10] Randic has shown how to translate this code into linear structural formulas.[11]

The algorithm utilized for generating the alkane series is similar to what one might use to list part of a dictionary from memory. First, try to extend the word base, then backtrack one letter and increment (eg do→doff→dog). This process can be shown as a binary tree diagram in which the left branch (word extension) is always less than the right branch (end-letter incrementation), but both branches are greater than the parent word base. With alkanes, a left branch is taken when one or more carbon atoms are removed from

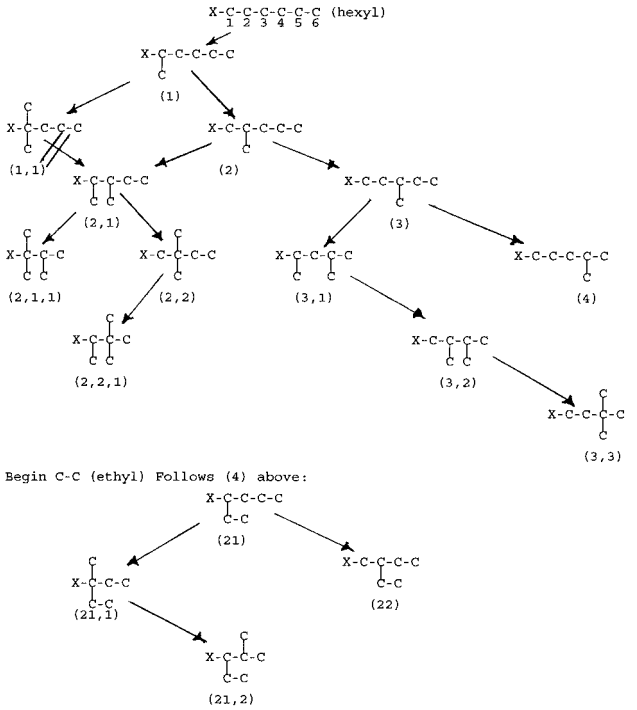

Figure 2. Tree diagram for the 17 hexyl isomers with ordered codes.

the chain end to form a side chain at the lowest locant permitted (2-methyl, 3-ethyl, etc.). This action is limited to situations where the shortening of the chain does not cause any other alkyl group to exceed its maximum locant, and the minimum locant is not doubly occupied. A right branch is taken when the lowest ranked alkyl group (by size, locant) is advanced to the next higher locant, provided that a vacancy exists and the move does not advance the group beyond its maximum locant. Since right branches increase isomer complexity more than left branches, they always follow left-branch attempts.

Before an alkane dictionary can be generated, it is necessary to rank the required alkyl groups by complexity to produce an "alphabet". Figure 2 shows the isomer tree for hexyl side chains. This is a simpler tree than those of the alkanes because all branches begin at locant 1, and there are no reverse-numbered duplicates to eliminate. The first two moves are left, giving a dimethyl isomer. While a right move now looks possible, this would disrupt the descending order of alkyl codes (here single digit numbers because the assigned rank for methyl groups is zero). Therefore, the third move is backward, in effect returning the second methyl group to the end of the chain. Now the lone methyl group can advance. (The chain end functions similarly to the third peg in the Tower of Hanoi puzzle that permits the transfer of the concentric disks from the first to the second peg without ever placing a larger disk atop a smaller one, as the rules require).

As the 12 methyl-only isomers are generated, the first methyl group moves down the chain until it reaches its maximum locant. Then 1-ethyl isomer appears, followed by two methyl isomers as left branches, and then the 2-ethyl isomer as a right branch. The tree growing process can also be simulated on a checkerboard, starting with six pieces in a row or column. The transition from 4-methyl to 1-ethyl can be regarded as the capture of the end piece by 4-methyl

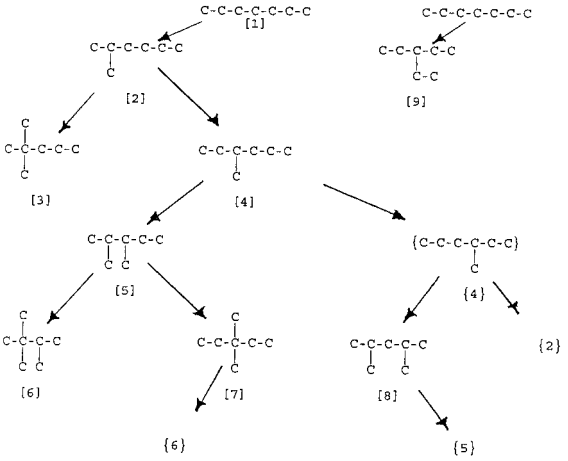

Figure 3. Heptane isomer tree showing reverse-numbered duplicates (braces).

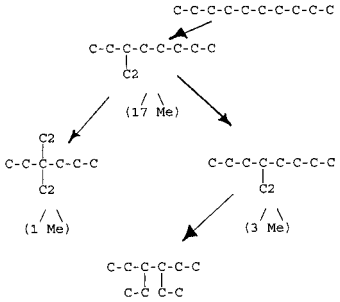

Figure 4. Ethyl isomers of decane tree showing nesting of methyl trees.

to become a king. The code for each isomer is obtained by appending a code for each alkyl group added. The numbers 1−20 are reserved for methyl locants. Larger alkyl groups are coded in the format rank/locant. Thus 1-ethyl has code 21, 2-propyl code 32 etc. Figure 3 shows the tree for heptane isomers (named and coded in Table 2). Here the tree includes reverse-numbered duplicates. These are easily identified during generation by tiebreaker tests. (Those that number the same both ways are labeled as symmetric main chains in the isomer tables.) Brackets label the duplicates in the diagram. Note that "4-methylhexane" is not a leaf node or dead end, but a necessary step on the path to 2,4-dimethylpentane. Figure 4 shows the nesting of the ethyl and methyl trees of decane (codes in Table 2).

There is an exception to the blocked advance of an alkyl group. If one (or more) of the next locants is doubly occupied by larger (or higher ranking) alkyl groups, the advancing group may leapfrog these if the landing locant has a vacancy and does not exceed the maximum. This feature is consistent with maintaining a treelike character by moving the smaller branches to the top of the tree. In terms of the developing alkyl codes, the leapfrog move is merely a descriptive fiction because the larger side chains are added first, so the smaller one is already past them to begin with (see C13-#649, 650 isomers at the top of Table 3).

BALANCING ACT

Before a given isomer can be confirmed, its leftmost and rightmost side chains must be checked to make sure that the end segments do not represent alternate side chains of lower complexity, indicating a duplicate structure. For example, a 2-ethyl group is illegal because this simplifies to a 3-methyl group on a longer chain. For C3+ alkyl groups, it is necessary to compare their rank with the end segment and its branches as a side chain. If the segment atom count is higher than that of the alkyl group, the configuration is legal, if lower illegal. Otherwise (if the same), segment small-branch detail must be evaluated. This process is facilitated by information already available in the developing isomer code. Simplified (2 or 3 digit) codes for the segment can be generated quickly for comparison with a fixed table for alkyl groups. For example, a 3-butyl group requires a 2,2-dimethyl segment to furnish the four atoms. This will also allow a *sec*-butyl or *tert*-butyl at the 3 position but not isobutyl which is more complex. Figure 5 shows the necessary left and right branch moves required to generate the four maximum complexity decane isomers (72−75 in Table 2). These are the tripropylmethanes with 0−3 isopropyl groups. Futile computer activity beyond that required to generate the four isomers is omitted. The number of moves required to generate the most complex decane isomer is not surprising, considering the effort generally necessary to be number one at anything. (Dozens of scratch paper doodlings that failed to give the complete tree sequence lacked an essential ingredient−patience.)

GENERATOR PROGRAM

Tables 2 and 3 show name-translated excerpts from program output. The descending numerical order of alkyl codes for each isomer is evident, as is the ascending (or dictionary) order of each set of codes from one isomer to the next. Both orderings provided valuable error checks during program development. Table 2 shows the 150 C1−C10 isomers, while Table 3 shows 154 of the most complex of the 802 C13 isomers. These include the five tetrapropylmethane isomers [$(C3)_4C$] that present 6-way main chain ties to the IUPAC Rules, and the 20 tributylmethanes [$(C4)_3C$] whose frequency distribution of the 4 C4 alkyl groups reflects their complexity. This is a more quantitative measure of complexity than rank. Table 4 gives distributions for C10−C21 alkanes and includes average numbers and sizes of alkyl groups.

The program to generate and name alkane isomers in alkyl-complexity order is written in about 400 lines of Fortran 77 and runs on a Sparc Ultra-4 Workstation. The only user input is the number of carbon atoms in the alkane. The current version generates all structural isomers of C1−21 and C22 except those with C7 side chains (Table 1). Output is in the form of isomer codes ready for name translation as shown in Tables 2 and 3. The program performs no matrix or sorting operations but works on the stack array of alkyl codes, as alkyl groups move down the chain, disappear, and then usually reemerge as atoms of a larger side chain.

Before comparing program performance with other generators, it is noted that most development efforts of the past decade have been directed toward expanding the variety of generated structures and user-supplied constraints. Programs designed to generate small subsets efficiently may not be optimal for exhaustive generation and vice-versa. Bohanec, in a 1995 paper in this journal, compares generation times for alkanes up to C16 with two other generators developed in years 1991−1994 with the GEN system.[12] The figures

Table 2. The First 150 Isomers (C1−C10) in Alkyl Group Complexity Order

				alkyl codes			
C1	1	methane	sym main				
C2	1	ethane	sym main				
C3	1	propane	sym main				
C4	1	butane	sym main				
C4	2	2-methylpropane	sym main	2			
C5	1	pentane	sym main				
C5	2	2-methylbutane		2			
C5	3	2,2-dimethylpropane	sym main	2	2		
C6	1	hexane	sym main				
C6	2	2-methylpentane		2			
C6	3	2,2-dimethylbutane		2	2		
C6	4	3-methylpentane	sym main	3			
C6	5	2,3-dimethylbutane	sym main	3	2		
C7	1	heptane	sym main				
C7	2	2-methylhexane		2			
C7	3	2,2-dimethylpentane		2	2		
C7	4	3-methylhexane		3			
C7	5	2,3-dimethylpentane		3	2		
C7	6	2,2,3-trimethylbutane		3	2	2	
C7	7	3,3-dimethylpentane	sym main	3	3		
C7	8	2,4-dimiethylpentane	sym main	4	2		
C7	9	3-ethylpentane	sym main	23			
C8	1	octane	sym main				
C8	2	2-methylheptane		2			
C8	3	2,2-dimethylhexane		2	2		
C8	4	3-methylheptane		3			
C8	5	2,3-dimethylhexane		3	2		
C8	6	2,2,3-trimethylpentane		3	2	2	
C8	7	3,3-dimethylhexane		3	3		
C8	8	2,3,3-trimethylpentane		3	3	2	
C8	9	2,2,3,3-tetramethylbutane	sym main	3	3	2	2
C8	10	4-methylheptane	sym main	4			
C8	11	2,4-dimethylhexane		4	2		
C8	12	2,2,4-trimethylpentane		4	2	2	
C8	13	3,4-dimethylhexane	sym main	4	3		
C8	14	2,3,4-trimethylpentane	sym main	4	3	2	
C8	15	2,5-dimethylhexane	sym main	5	2		
C8	16	3-ethylhexane		23			
C8	17	2-methyl-3-ethylpentane		23	2		
C8	18	3-methyl-3-ethylpentane	sym main	23	3		
C9	1	nonane	sym main				
C9	2	2-methyloctane		2			
C9	3	2,2-dimethylheptane		2	2		
C9	4	3-methyloctane		3			
C9	5	2,3-dimethylheptane		3	2		
C9	6	2,2,3-trimethylhexane		3	2	2	
C9	7	3,3-dimethylheptane		3	3		
C9	8	2,3,3-trimethylhexane		3	3	2	
C9	9	2,2,3,3-tetramethylpentane		3	3	2	2
C9	10	4-methyloctane		4			
C9	11	2,4-dimethylheptane		4	2		
C9	12	2,2,4-trimethylhexane		4	2	2	
C9	13	3,4-dimethylheptane		4	3		
C9	14	2,3,4-trimethylhexane		4	3	2	
C9	15	2,2,3,4-tetramethylpentane		4	3	2	2
C9	16	3,3,4-trimethylhexane		4	3	3	
C9	17	2,3,3,4-tetramethylpentane	sym main	4	3	3	2
C9	18	4,4-dlmethylheptane	sym main	4	4		
C9	19	2,4,4-trimethylhexane		4	4	2	
C9	20	2,2,4,4-tetramethylpentane	sym main	4	4	2	2
C9	21	2,5-dlmethylheptane		5	2		
C9	22	2,2,5-trimethylhexane		5	2	2	
C9	23	3,5-dimethylheptane	sym main	5	3		
C9	24	2,3,5-trimethylhexane		5	3	2	
C9	25	2,6-dimethylheptane	sym main	6	2		
C9	26	3-ethylheptane		23			
C9	27	2-methyl-3-ethylhexane		23	2		
C9	28	2,2-dimethyl-3-ethylpentane		23	2	2	
C9	29	3-methyl-3-ethylhexane		23	3		
C9	30	2,3-dimethyl-3-ethylpentane		23	3	2	
C9	31	4-methyl-3-ethylhexane		23	4		
C9	32	2,4-dimethyl-3-ethylpentane	sym main	23	4	2	
C9	33	5-methyl-3-ethylhexane		23	5		
C9	34	3,3-diethylpentane	sym main	23	23		
C9	35	4-ethylheptane	sym main	24			

Table 2 (Continued)

				alkyl codes				
C10	1	decane	sym main					
C10	2	2-methylnonane		2				
C10	3	2,2-dimethyloctane		2	2			
C10	4	3-methylnonane		3				
C10	5	2,3-dimethyloctane		3	2			
C10	6	2,2,3-trimethylheptane		3	2	2		
C10	7	3,3-dimethyloctane		3	3			
C10	8	2,3,3-trimethylheptane		3	3	2		
C10	9	2,2,3,3-tetramethylhexane		3	3	2	2	
C10	10	4-methylnonane		4				
C10	11	2,4-dimethyloctane		4	2			
C10	12	2,2,4-trimethylheptane		4	2	2		
C10	13	3,4-dimethyloctane		4	3			
C10	14	2,3,4-trimethylheptane		4	3	2		
C10	15	2,2,3,4-tetramethylhexane		4	3	2	2	
C10	16	3,3,4-trimethylheptane		4	3	3		
C10	17	2,3,3,4-tetramethylhexane		4	3	3	2	
C10	18	2,2,3,3,4-pentamethylpentane		4	3	3	2	2
C10	19	4,4-dimethyloctane		4	4			
C10	20	2,4,4-trimethylheptane		4	4	2		
C10	21	2,2,4,4-tetramethylhexane		4	4	2	2	
C10	22	3,4,4-trimethylheptane		4	4	3		
C10	23	2,3,4,4-tetramethylhexane		4	4	3	2	
C10	24	2,2,3,4,4-pentamethylpentane	sym main	4	4	3	2	2
C10	25	3,3,4 4-tetramethylhexane	sym main	4	4	3	3	
C10	26	5-methylnonane	sym main	5				
C10	27	2,5-dimethyloctane		5	2			
C10	28	2,2,5-trimethylheptane		5	2	2		
C10	29	3,5-dimethyloctane		5	3			
C10	30	2,3,5-trimethylheptane		5	3	2		
C10	31	2,2,3,5-tetramethylhexane		5	3	2	2	
C10	32	3,3,5-trimethylheptane		5	3	3		
C10	33	2,3,3,5-tetramethylhexane		5	3	3	2	
C10	34	4,5-dimethyloctane	sym main	5	4			
C10	35	2,4,5-trimethylheptane		5	4	2		
C10	36	2,2,4,5-tetramethylhexane		5	4	2	2	
C10	37	3,4,5-trimethylheptane	sym main	5	4	3		
C10	38	2,3,4,5-tetramethylhexane	sym main	5	4	3	2	
C10	39	2,5,5-trimethylheptane		5	5	2		
C10	40	2,2,5,5-tetramethylhexane	sym main	5	5	2	2	
C10	41	2,6-dimethyloctane		6	2			
C10	42	2,2,6-trimethylheptane		6	2	2		
C10	43	3,6-dimethyloctane	sym main	6	3			
C10	44	2,3,6-trimethylheptane		6	3	2		
C10	45	2,4,6-trimethylheptane	sym main	6	4	2		
C10	46	2,7-dimethyloctane	sym main	7	2			
C10	47	3-ethyloctane		23				
C10	48	2-methyl-3-ethylheptane		23	2			
C10	49	2,2-dimethyl-3-ethylhexane		23	2	2		
C10	50	3-methyl-3-ethylheptane		23	3			
C10	51	2,3-dimethyl-3-ethylhexane		23	3	2		
C10	52	2,2,3-trimethyl-3-ethylpentane		23	3	2	2	
C10	53	4-methyl-3-ethylheptane		23	4			
C10	54	2,4-dimethyl-3-ethylhexane		23	4	2		
C10	55	2,2,4-trimethyl-3-ethylpentane		23	4	2	2	
C10	56	3,4-dimethyl-3-ethylhexane		23	4	3		
C10	57	2,3,4-trimethyl-3-ethylpentane	sym main	23	4	3	2	
C10	58	4,4-dimethyl-3-ethylhexane		23	4	4		
C10	59	5-methyl-3-ethylheptane		23	5			
C10	60	2,5-dimethyl-3-ethylhexane		23	5	2		
C10	61	3,5-dimethyl-3-ethylhexane		23	5	3		
C10	62	4,5-dimethyl-3-ethylhexane		23	5	4		
C10	63	5,5-dimethyl-3-ethylhexane		23	5	5		
C10	64	6-methyl-3-ethylheptane		23	6			
C10	65	3,3-diethylhexane		23	23			
C10	66	2-methyl-3,3-diethylpentane		23	23	2		
C10	67	4-ethyloctane		24				
C10	68	2-methyl-4-ethylheptane		24	2			
C10	69	3-methyl-4-ethylheptane		24	3			
C10	70	4-methyl-4-ethylheptane	sym main	24	4			
C10	71	3,4-diethylhexane	sym main	24	23			
C10	72	2-methyl-3-propylhexane	(1st non−IUPAC main chain)	33	2			
C10	73	2,4-dimethyl-3-propylpentane	(2nd) sym main	33	4	2		
C10	74	4-propylheptane	sym main	34				
C10	75	2,4-dlmethyl-3-isopropylpentane	sym main	43	4	2		

Table 3. Tridecane Isomers, Beginning with the First Tetra_C3 Methane

C13#	alkyl complexity-order name	feature	alkyl codes					
649	2-methyl-3,3-dipropylhexane	triPr isoPrM	33	33	2			
650	2,4-dimethyl-3,3-dipropylpentane	diPr diisoPrM sym	33	33	4	2		
651	4-propyldecane		34					
652	2-methyl-4-propylnonane		34	2				
653	2,2-dimethyl-4-propyloctane		34	2	2			
654	3-methyl-4- propylnonane		34	3				
655	2,3-dimethyl-4-propyloctane		34	3	2			
656	2,2,3-trimethyl-4-propylheptane		34	3	2	2		
657	3,3-dimethyl-4-propyloctane		34	3	3			
658	2,3,3-trimethyl-4-propylheptane		34	3	3	2		
659	4-methyl-4-propylnonane		34	4				
660	2,4-dimethyl-4-propyloctane		34	4	2			
661	2,2,4-trimethyl-4-propylheptane		34	4	2	2		
662	3,4-dimethyl-4-propyloctane		34	4	3			
663	2,3,4-trimethyl-4-propylheptane		34	4	3	2		
664	3,3,4-trimethyl-4-propylheptane		34	4	3	3		
665	5-methyl-4-propylnonane		34	5				
666	2,5-dimethyl-4-propyloctane		34	5	2			
667	2,2,5-trimethyl-4-propylheptane		34	5	2	2		
668	3,5-dimethyl-4-propyloctane		34	5	3			
669	2,3,5-trimethyl-4-propylheptane		34	5	3	2		
670	3,3,5-trimethyl-4-propylheptane		34	5	3	3		
671	4,5-dimethyl-4-propyloctane		34	5	4			
672	2,4,5-trimethyl-4-propylheptane		34	5	4	2		
673	3,4,5-trimethyl-4-propylheptane	sym main	34	5	4	3		
674	5,5-dimethyl-4-propyloctane		34	5	5			
675	2,5,5-trimethyl-4-propylheptane		34	5	5	2		
676	6-methyl-4-propylnonane		34	6				
677	2,6-dimethyl-4-propyloctane		34	6	2			
678	2,2,6-trimethyl-4-propylheptane		34	6	2	2		
679	3,6-dimethyl-4-propyloctane		34	6	3			
680	2,3,6-trimethyl-4-propylheptane		34	6	3	2		
681	4,6-dimethyl-4-propyloctane		34	6	4			
682	2,4,6-trimethyl-4-propylheptane	sym main	34	6	4	2		
683	5,6-dimethyl-4-propyloctane		34	6	5			
684	6,6-dimethyl-4-propyloctane		34	6	6			
685	7-methyl-4-propylnonane		34	7				
686	2,7-dimethyl-4-propyloctane		34	7	2			
687	3,7-dimethyl-4-propyloctane		34	7	3			
688	4,7-dimethyl-4-propyloctane		34	7	4			
689	5,7-dimethyl-4-propyloctane		34	7	5			
690	6,7-dimethyl-4-propyloctane		34	7	6			
691	7,7-dimethyl-4-propyloctane		34	7	7			
692	8-methyl-4-propylnonane		34	8				
693	3-ethyl-4-propyloctane		34	23				
694	2-methyl-3-ethyl-4-propylheptane		34	23	2			
695	3-methyl-3-ethyl-4-propylheptane		34	23	3			
696	4-methyl-3-ethyl-4-propylheptane		34	23	4			
697	5-methyl-3-ethyl-4-propylheptane		34	23	5			
698	6-methyl-3-ethyl-4-propylheptane		34	23	6			
699	4-ethyl-4-propyloctane		34	24				
700	2-methyl-4 -ethyl-4-propylheptane		34	24	2			
701	3-methyl-4-ethyl-4-propylheptane		34	24	3			
702	5-ethyl-4-propyloctane		34	25				
703	6-ethyl-4-propyloctane		34	26				
704	4,4-dipropylheptane	tetraPrM sym main	34	34				
705	5-propyldecane		35					
706	2-methyl-5-propylnonane		35	2				
707	3-methyl-5-propylnonane		35	3				
708	4-methyl-5-propylnonane		35	4				
709	5-methyl-5-propylnonane	sym main	35	5				
710	2-methyl-3-isopropylnonane		43	2				
711	2,2-dimethyl-3-isopropyloctane		43	2	2			
712	2,3-dimethyl-3-isopropyloctane		43	3	2			
713	2,2,3-trimethyl-3-isopropylheptane		43	3	2	2		
714	2,4-dimethyl-3-isopropyloctane		43	4	2			
715	2,2,4-trimethyl-3-isopropylheptane		43	4	2	2		
716	2,3,4-trimethyl-3-isopropylheptane		43	4	3	2		
717	2,2,3,4-tetramethyl-3-isopropylhexane		43	4	3	2	2	
718	2,4,4-trimethyl-3-isopropylheptane		43	4	4	2		
719	2,2,4,4-tetramethyl-3-isopropylhexane		43	4	4	2	2	
720	2,3,4,4-tetramethyl-3-isopropylhexane		43	4	4	3	2	
721	2,2,3,4,4-pentamethyl-3-isopropylpentane	sym main	43	4	4	3	2	2
722	2,5-dimethyl-3-isopropyloctane		43	5	2			
723	2,2,5-trimethyl-3-lsopropylheptane		43	5	2	2		
724	2,3,5-trimethyl-3-isopropylheptane		43	5	3	2		
725	2,2,3,5-tetramethyl-3-isopropylhexane		43	5	3	2	2	

Table 3 (Continued)

C13#	alkyl complexity-order name	feature	alkyl codes				
726	2,4,5-trimethyl-3-isopropylheptane		43	5	4	2	
727	2,2,4,5-tetramethyl-3-isopropylhexane		43	5	4	2	2
728	2,3,4,5-tetramethyl-3-isopropylhexane		43	5	4	3	2
729	2,4,4,5-tetramethyl-3-isopropylhexane		43	5	4	4	2
730	2,5,5-trimethyl-3-isopropylheptane		43	5	5	2	
731	2,2,5,5-tetramethyl-3-isopropylhexane		43	5	5	2	2
732	2,3,5,5-tetramethyl-3-isopropylhexane		43	5	5	3	2
733	2,4,5,5-tetramethyl-3-isopropylhexane		43	5	5	4	2
734	2,6-dimethyl-3-isopropyloctane		43	6	2		
735	2,2,6-trimethyl-3-isopropylheptane		43	6	2	2	
736	2,3,6-trimethyl-3-isopropylheptane		43	6	3	2	
737	2,4,6-trimethyl-3-isopropylheptane		43	6	4	2	
738	2,5,6-trimethyl-3-isopropylheptane		43	6	5	2	
739	2,6,6-trimethyl-3-isopropylheptane		43	6	6	2	
740	2,7-dimethyl-3-isopropyloctane		43	7	2		
741	2-methyl-3-ethyl-3-isopropylheptane		43	23	2		
742	2,4-dimethyl-3-ethyl-3-isopropylhexane		43	23	4	2	
743	2,2,4-trimethyl-3-ethyl-3-isopropylpentane		43	23	4	2	2
744	2,5-dimethyl-3-ethyl-3-isopropylhexane		43	23	5	2	
745	2-methyl-4-ethyl-3-isopropylheptane		43	24	2		
746	2,2-dimethyl-4-ethyl-3-isopropylhexane		43	24	2	2	
747	2,3-dimethyl-4-ethyl-3-isopropylhexane		43	24	3	2	
748	2,4-dimethyl-4-ethyl-3-isopropylhexane		43	24	4	2	
749	2,5-dimethyl-4-ethyl-3-isopropylhexane		43	24	5	2	
750	2-methyl-5-ethyl-3-isopropylheptane		43	25	2		
751	2,4-dimethyl-3-propyl-3-isopropylpentane	Pr triisoPrM	43	33	4	2	
752	2,4-dimethyl-3,3-diisopropylpentane	tetraisoPrM sym	43	43	4	2	
753	2-methyl-4-isopropylnonane		44	2			
754	2,2-dimethyl-4-isopropyloctane		44	2	2		
755	3-methyl-4-isopropylnonane		44	3			
756	2,3-dimethyl-4-isopropyloctane		44	3	2		
757	3,3-dimethyl-4-isopropyloctane		44	3	3		
758	2,4-dimethyl-4-isopropyloctane		44	4	2		
759	3,4-dimethyl-4-isopropyloctane		44	4	3		
760	2,5-dimethyl-4-isopropyloctane		44	5	2		
761	2,2,5-trimethyl-4-isopropylheptane		44	5	2	2	
762	3,5-dimethy1-4-isopropyloctane		44	5	3		
763	2,3,5-trimethyl-4-isopropylheptane		44	5	3	2	
764	3,3,5-trimethyl1-4-isopropylheptane		44	5	3	3	
765	2,4,5-trimethyl-4-isopropylheptane		44	5	4	2	
766	3,4,5-trimethyl-4-isopropylheptane	sym main	44	5	4	3	
767	2,5,5-trimethyl-4-isopropylheptane		44	5	5	2	
768	2,6-dimethyl-4-isopropyloctane		44	6	2		
769	2,2,6-trimethyl-4-isopropylheptane		44	6	2	2	
770	3,6-dimethy1-4-isopropyloctane		44	6	3		
771	2,3,6-trimethyl-4-lsopropylheptane		44	6	3	2	
772	2,4,6-trimethyl-4-isopropylheptane	sym main	44	6	4	2	
773	2,7-dimethyl-4-isopropyloctane		44	7	2		
774	3,7-dimethyl-4-isopropyloctane		44	7	3		
775	3-ethyl-4-isopropyloctane		44	23			
776	5-methyl-3-ethyl-4-isopropylheptane		44	23	5		
777	6-methyl-3-ethyl-4-isopropylheptane		44	23	6		
778	5-isopropyldecane		45				
779	2-methyl-5-isopropylnonane		45	2			
780	3-methyl-5-isopropylnonane		45	3			
781	4-methyl-5-isopropylnonane		45	4			
782	5-methyl-5-isopropylnonane	sym main	45	5			
783	2,2-dimethyl-3-butylheptane	--20 x-butyl isomers--	53	2	2		
784	2,2,4-trimethyl-3-butylhexane		53	4	2	2	
785	2,2,4,4-tetramethyl-3-butylpentane	sym main	53	4	4	2	2
786	2,2,5-trimethyl-3-butylhexane		53	5	2	2	
787	2-methyl-4-butyloctane		54	2			
788	3-methyl-4-butyloctane	10 butyl	54	3			
789	2,5-dimethyl-4-butylheptane		54	5	2		
790	3,5-dimethyl-4-butylheptane	sym main	54	5	3		
791	2,6-dimethyl-4-butylheptane	sym main	54	6	2		
792	5-butylnonane	sym main	55				
793	2,2,4-trimethyl-3-sec-butylhexane		63	4	2	2	
794	2,2,4,4-tetramethyl-3-sec-butylpentane	sym main	63	4	4	2	2
795	2,2,5-trimethyl-3-sec-butylhexane	6 sec-butyl	63	5	2	2	
796	2,5-dimethyl-4-sec-butylheptane		64	5	2		
797	3,5-dimethyl-4-sec-butylheptane	sym main	64	5	3		
798	2,6-dimethy1-4-sec-butylheptane	sym main	64	6	2		
799	2,2,4,4-tetramethyl-3-tert-butylpentane	sym main	73	4	4	2	2
800	2,2,5-trimethyl-3-tert-butylhexane	3 tert-butyl	73	5	2	2	
801	2,6-dimethyl-4-tert-butylheptane	sym main	74	6	2		
802	2,6-dimethyl-4-isobutylheptane	1 isobutyl sym	84	6	2		

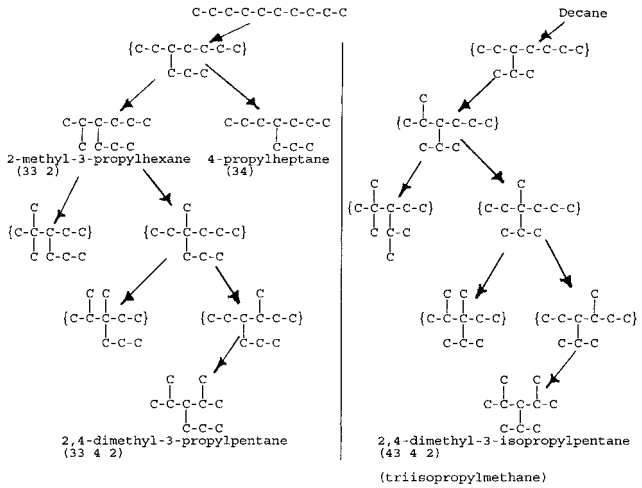

Figure 5. Tree diagram and isomer codes for C10 tripropylmethanes.

given are 372 s for 1858 C14 and 18046 s for 10359 C16. The tests were done on a 486 PC (33 MHz). Allowing a 100-fold speed increase for today's faster computers translates to about 500/s for C14 and 60/s for C16. Table 5 shows the generation rate here for C10−C22 in millions of isomer codes per second, exclusive of output and name translation. The corresponding rates are seen to be about 800 000 for C14 and 700 000 for C16. The rate is nearly constant for each set of three alkanes having the same largest alkyl group and is inverse linear with size for each multiple of three. While an alkane-only generator is expected to be faster than a multiple structure generator, the high intrinsic speed and its much lower decrease with alkane size encourage pursuit of extension to other structure types.

Although C19 isomers with C6 side chains comprise less than 1% of the total, their generation consumes about 15% of the run time. This is because they are the most complex and require increasingly more structure balancing as maximum complexity is approached (see Figure 5). The percentage of C6 isomers in C20 is nearly three times higher, but the generation rate is about the same. Addition of one carbon atom reduces complexity for most isomers and allows many to be verified by chain-end fragment size alone. While generation of high complexity isomers is harder, enumeration is easier because there are fewer configurations to consider. For example, for C20 the number of C6 isomers is calculated as below, providing an independent check on program results:

$$\#\,\mathrm{C7{-}C{-}(C6)2} + \#\,\mathrm{(C6)3\text{-}C{-}C}$$

$$= 39\mathrm{C}(18,2) + \mathrm{C}(19,3)$$

$$= 5967 + 969$$

$$= 6936\ (\text{sum of C6 in Table 4})$$

Incremental gains in speed have been obtained through improved, simple screening tests that reduce the number of function calls that examine all the side chains. For example, when the highest ranking side chain is less than halfway across the main chain, the reflective symmetry that might otherwise create reverse-numbered duplicates is disrupted and all legal combinations of lower-ranked side chains are valid (see Table 2 C10-#47). There were three modifications that produced major (>20%) gains in speed. The first resulted as a byproduct of changing the two-digit display alkyl codes (rank, locant) from base 10 to base 64 to extend their range. This reduced extraction of the rank and locant separately to shift and logical-and instructions. Shortly thereafter, the (nonstandard) direct recursion of alkyl ranks was eliminated, because the alkyl codes array is common to all alkyl groups, and when the first of a next higher-ranking alkyl group is added, the recursion goes nonstop to the bottom to add methyl groups, similar to adding an "a" after starting the next higher letter in a dictionary. Finally, it was found that a dead end is reached whenever a left-branch tree move produces a reverse-numbered duplicate, allowing immediate backtracking.

ULTIMATE COMPACT CODE

The alkane isomer codes (size + alkyl codes) shown in Tables 2 and 3 comprise a reversible code that readily translates to a complexity-ordered name when read from right to left. (Since the codes completely specify structures, translation to IUPAC names, substituted methane names or any other structure-based name is also enabled.) Utilizing methods similar to those employed in ref 2, alkane codes up to about C20 can be compressed to 32-bit positive integers. Here the main chain length would not follow the highest order descriptor (size) as required by the IUPAC rules but would be obtained by difference after specifying all the side chains. The code for the highest ranking alkyl group and its locant would follow the size. The (N-1)/3 formula would be applied to the remaining carbon atoms to obtain a maximum size for the next side chain. Unfortunately, this is offset by the initially large locant range for which bits must be reserved because the main chain length is unknown.

On the other hand, if the isomer code of an N-carbon alkane is determined from its structure manually or by computer from a connection table, its unique sequence number starting from methane = 1 can be obtained from the generator program by matching and then adding the literature values for isomer counts from C1 to C(N-1). This is the ultimate code in terms of compactness because there is no wasted space. Alternatively, the isomer codes can be stored in a direct access file, so that the matching sequence number is just the record number of the file for decoding, and a simple binary search for an isomer code will quickly find its sequence number for encoding. A compromise 32-bit integer code could use the left half to store the size and one or two highest ranked alkyl groups and their locants and the right half to store the sequence number offset from a table. The generator would initialize the codes array from the left half and then run until the number of isomers generated from there matched the offset in the right half. This would be analogous to opening a dictionary to the first two or three letters of a sought word and then visually scanning from there.

The current version of the generator program can also code and name alkanes larger than C22 up to the point where C7 alkyls are required. As a final example that illustrates most of the generating algorithm properties, ref 11 displays a C23 complex alkane (structure 10) that is numbered and named there according to nodal nomenclature rules. Manual coding and computer processing by minimum side-chain complexity

Table 4. Alkyl Frequency Distribution by Complexity for C10−C21 Isomers

	total	methyl	ethyl	propyl	isoPr	butyl	sec-Bu	tert-Bu	isoBu
C10	203	171	28	3	1	0	0	0	0
percent	100	84.24	13.79	1.48	0.49	0.00	0.00	0.00	0.00
Isomers 75; Averages: Alkyls per Isomer 2.71; Alkyl Size 1.18									
	total	methyl	ethyl	propyl	isoPr	butyl	sec-Bu	ter-Bu	isoBu
C11	488	398	74	11	5	0	0	0	0
percent	100	81.56	15.16	2.25	1.02	0.00	0.00	0.00	0.00
Isomers 159; Averages: Alkyls per Isomer 3.07; Alkyl Size 1.22									
	total	methyl	ethyl	propyl	isoPr	butyl	sec-Bu	ter-Bu	isoBu
C12	1211	962	189	37	23	0	0	0	0
percent	100	79.44	15.61	3.06	1.90	0.00	0.00	0.00	0.00
Isomers 355; Averages: Alkyls per Isomer; 3.41 Alkyl Size 1.26									
	total	methyl	ethyl	propyl	isoPr	butyl	sec-Bu	ter-Bu	isoBu
C13	3004	2323	480	107	74	10	6	3	1
percent	100	77.33	15.98	3.56	2.46	0.33	0.20	0.10	0.03
Isomers 802; Averages: Alkyls per Isomer 3.75; Alkyl Size 1.30									
	total	methyl	ethyl	propyl	isoPr	butyl	sec-Bu	ter-Bu	isoBu
C14	7565	5720	1226	296	223	42	30	19	9
percent	100	75.61	16.21	3.91	2.95	0.56	0.40	0.25	0.12
Isomers 1858; Averages: Alkyls per Isomer 4.07; Alkyl Size 1.34									
	total	methyl	ethyl	propyl	isoPr	butyl	sec-Bu	ter-Bu	isoBu
C15	19058	14090	3136	790	628	146	117	89	62
percent	100	73.93	16.46	4.15	3.30	0.77	0.61	0.47	0.33
Isomers 4347; Averages: Alkyls per Isomer 4.38; Alkyl Size 1.38									
	total	methyl	ethyl	propyl	isoPr	butyl	sec-Bu	ter-Bu	isoBu
C16	48615	35229	8062	2106	1730	448	376	306	238
percent	100	72.47	16.58	4.33	3.56	0.92	0.77	0.63	0.49
	pentyl	1-MeBu	ter-Pe	2-MeBu	1,2MPr	neoPe	3-MeBu	1-EtPr	
C16 (cont'd)	36	28	21	15	10	6	3	1	
percent	0.07	0.06	0.04	0.03	0.02	0.01	0.01	0.00	
Isomers 10359; Averages: Alkyls per Isomer 4.69; Alkyl Size 1.42									
	total	methyl	ethyl	propyl	isoPr	butyl	sec-Bu	ter-Bu	isoBu
C17	124308	88365	20782	5545	4656	1311	1142	972	803
percent	100	71.09	16.72	4.46	3.75	1.05	0.92	0.78	0.65
	pentyl	1-MeBu	ter-Pe	2-MeBu	1,2MPr	neoPe	3-MeBu	1-EtPr	
C17 (cont'd)	172	147	123	100	78	57	37	18	
percent	0.14	0.12	0.10	0.08	0.06	0.05	0.03	0.01	
Isomers 24894; Averages: Alkyls per Isomer 4.99; Alkyl Size 1.46									
	total	methyl	ethyl	propyl	isoPr	butyl	sec-Bu	ter-Bu	isoBu
C18	320070	223613	53720	14634	12491	3656	3266	2868	2462
percent	100	69.86	16.78	4.57	3.90	1.14	1.02	0.90	0.77
	pentyl	1-MeBu	ter-Pe	2-MeBu	1,2MPr	neoPe	3-MeBu	1-EtPr	
C18 (cont'd)	637	573	510	448	387	327	268	210	
percent	0.20	0.18	0.16	0.14	0.12	0.10	0.08	0.07	
Isomers 60523; Averages Alkyls per Isomer 5.29; Alkyl Size 1.49									
	total	methyl	ethyl	propyl	isoPr	butyl	sec-Bu	ter-Bu	isoBu
C19	826087	568352	139306	38524	33285	10055	9135	8198	7244
percent	100	68.80	16.86	4.66	4.03	1.22	1.11	0.99	0.88
	pentyl	1-MeBu	ter-Pe	2-MeBu	1,2MPr	neoPe	3-MeBu	1-EtPr	all-C6
C19 (cont'd)	2048	1887	1728	1571	1416	1263	1112	963	969
percent	0.25	0.23	0.21	0.19	0.17	0.15	0.13	0.12	0.12
Isomers 148284; Averages: Alkyls per Isomer 5.57; Alkyl Size 1.53									
	total	methyl	ethyl	propyl	isoPr	butyl	sec-Bu	ter-Bu	isoBu
C20	2140016	1452689	362511	101628	88646	27302	25087	22833	20540
percent	100	67.88	16.94	4.75	4.14	1.28	1.17	1.07	0.96
	pentyl	1-MeBu	ter-Pe	2-MeBu	1,2MPr	neoPe	3-MeBu	1-EtPr	all-C6
C20 (cont'd)	6237	5832	5431	5034	4641	4252	3867	3486	6936
percent	0.29	0.27	0.25	0.24	0.22	0.20	0.18	0.16	0.32
Isomers 366319; Averages: Alkyls per Isomer; 5.84 Alkyl Size 1.55									

Table 4 (Continued)

	total	methyl	ethyl	propyl	isoPr	butyl	sec-Bu	ter-Bu	isoBu
C21	5557249	3729433	944281	268228	235751	73705	68299	62804	57220
percent	100	67.11	16.99	4.83	4.24	1.33	1.23	1.13	1.03
	pentyl	1-MeBu	ter-Pe	2-MeBu	1,2MPr	neoPe	3-MeBu	1-EtPr	all-C6
C21 (cont'd)	18100	17131	16157	15180	14202	13225	12251	11282	33813
percent	0.33	0.31	0.29	0.27	0.26	0.24	0.22	0.20	0.61

Isomers 910726; Averages: Alkyls per Isomer 6.10; Alkyl Size 1.58

Table 5. Rates of Computer Generation (Integer Codes) for C10−C22 Alkanes

alkane C##	structural isomers	max alkyl	#/sec (mil)	rate x ##
10	75	\	1.042	
11	159	C3	1.026	
12	355	/	1.014	12.2
13	802	\	0.802	
14	1828	C4	0.801	
15	4347	/	0.799	12.0
16	10359	\	0.700	
17	24284	C5	0.692	
18	60523	/	0.683	12.3
19	148284	\	0.598	
20	366319	C6	0.586	
21	910726	/	0.580	12.2
22 (-C7)	2267998	C7	0.579	

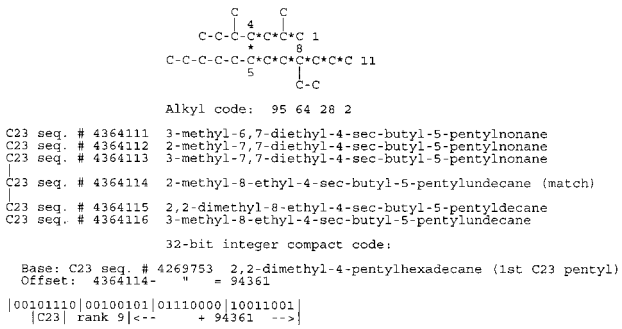

Figure 6. Unique sequence number and compact code for a C23 complex alkane.

is shown here in Figure 6: First, simplify the C9 side chain by routing the main chain through one of the two C4 (butyl) groups. Choose the isobutyl group because it is the more complex. Complete the chain so as to include 8-ethyl. Now enter the isomer codes (as in Table 3, + 95 for 5-pentyl) and run the generator. In about 6 s of CPU time the code is matched, producing the unique sequence number. (With about 5.7 million C23 structural isomers, the complexity rank of this alkane is in the 76th percentile.) Its three precursors illustrate the final generating steps: First, 5-pentyl is central to the nonane main chain, so 4-*sec*-butyl controls its numbering. 7,7-Diethyl is the last diethyl configuration generated, while methyl is held captive at the 2 and 3 positions to balance *sec*-butyl. Finally, 3-methyl and 7-ethyl are returned to the end of the chain by backtracking, the remaining ethyl advances to locant 8, and methyl is recycled to locant 2 to balance 4-*sec*-butyl.

A simple, 30-bit complexity ordered compact code (Figure 6) is obtained by storing 23 in the first 5 bits, 9—the rank of the most complex side chain (C5) in the next 7 (allowing space for 33 C1−C6 and 39 C7) and 18 for the offset from the sequence number of the first isomer containing C5. To retrieve the isomer name from the code, the size and rank are extracted from the code, and the alkyl codes array is initialized with 94 (rank, min locant). With the sequence number of the first isomer of highest rank taken from a table, the generator only has to count isomers up to the offset value of 94361 instead of 4.36 million to obtain the complete isomer code and the name.

CONCLUSION

A new method for generating the alkane series in a manner that concurrently imposes a dictionary ordering has been presented. It is hoped that this method can be extended to alkenes, rings, and other more complex structures as the N_tuple codes have been over the years.

REFERENCES AND NOTES

(1) Davidson, S. An Improved IUPAC−Based Method for Identifying Alkanes. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 151−155.
(2) Davidson, S. Compact Numeric Alkane Codes Derived from IUPAC Nomenclature. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 417−422.
(3) Davidson, S. Algorithm for Selecting the Parent Structural Unit of a Ring-Chain Assembly. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 215−221.
(4) IUPAC. *Nomenclature of Organic Compounds*; Butterworth: London, 1958; pp 8−9.
(5) IUPAC. *Nomenclature of Organic Compounds;* Pergamon Press: New York, 1979; pp 10−11.
(6) Henze, H. R.; Blair, C. The Number of Isomeric Hydrocarbons of the Methane Series. *J. Am. Chem. Soc.* **1931**, *53*, 3077−3085.
(7) Bytautas, L.; Klein, D. J. Chemical Combinatorics of Alkane-Isomer Enumeration and More. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1063−1078.
(8) Roberts, J. D. Caltech, 1954 (personal recollection).
(9) Knop, J. V.; Muller, W. R.; Jericevic, Z.; Trinajstic, N. Computer Enumeration and Generation of Trees and Rooted Trees. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 91−99.
(10) Contreras, M. L.; Alvarez, J.; Riveros, M.; Arias, G.; Rozas, R. Exhaustive Generation of Organic Isomers. 6. Stereoisomers Having Isolated and Spiro Cycles and New Extended N_tuples. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 964−977.
(11) Randic, M.; Nikolic, S.; Trinajstic, N. Compact Codes: On Nomenclature of Acyclic Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 357−365.
(12) Bohanec, S. Structure Generation by the Combination of Structure Reduction and Structure Assembly. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 494−503.

CI010094B