analytical chemistry

Article

pubs.acs.org/ac

Isobaric Molecular Formulae of C, H, and O: A View from the Negative Quadrants of van Krevelen Space

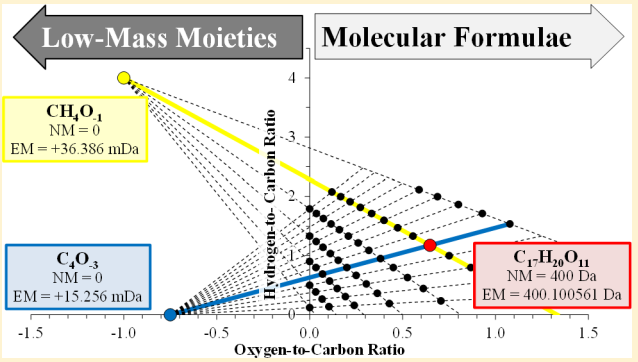
E. Michael Perdue*,† and Nelson W. Green†,‡

†Department of Chemistry, Ball State University, Muncie, Indiana 47306, United States

‡School of Chemical and Biomolecular Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332-0340, United States

Supporting Information

ABSTRACT: When compositions of an isobaric series of molecular formulae consisting of C, H, and O are displayed in a van Krevelen plot, a remarkable number of nonparallel lines of compositions can be observed. Each group of related lines converge at a point in a negative quadrant of van Krevelen space (e.g., H/C = 4, O/C = −1). These points of convergence have H/C and O/C ratios that correspond to "molecular formulae" in which the stoichiometric coefficients of some of the isotopes are negative. In this manner, a group of related low-mass moieties have been identified (CH_4O_{-1}, C_4O_{-3}, $C_2H_{-8}O_{-1}$, CH_{-12}, etc.). Each of these moieties has a nominal mass of zero and a very small exact mass. Furthermore, all of these low-mass moieties have compositions that fall on the line H/C = −12 − 16(O/C), which lies entirely in the negative quadrants of van Krevelen space. This paper demonstrates that all low-mass moieties consisting of C, H, and O can be expressed formally as linear combinations of any two moieties. Likewise, all molecular formulae that fall on a line that passes through a given low-mass moiety must differ compositionally by a multiple of the composition of that moiety, and their exact masses must differ by a multiple of the exact mass of the moiety. This latter relationship has been invoked for very rapid assignment of a molecular formula to an exact mass. Finally, a more comprehensive and theoretically based understanding of family scores has been developed around the concept of low-mass moieties.



van Krevelen[1] introduced his eponymous plot of atomic H/C versus atomic O/C as a means of visualizing diagenetic and catagenetic transformations by which biomass is eventually converted into coal. The van Krevelen plot was shown to be useful for displaying chemical compositional data, for distinguishing between structural classes of molecules, and for describing chemical reactions of organic matter in terms of gains/losses of small volatile compounds such as CH_4, CO_2, H_2O, H_2, and O_2. This plot was adopted to describe chemical compositions and reactions of kerogens in sedimentary rocks,[2] soil humic acids,[3] sedimentary organic matter,[4] and aquatic humic substances.[5] The paper by Reuter and Perdue[4] is noteworthy because it demonstrates that diagenetic changes of organic matter can be depicted in a van Krevelen plot as gains/losses of lipids, sugars, and proteins. Sun et al.[6] extended this use of the van Krevelen plot to describe diagenetic alteration of dissolved organic matter in a Georgia river. Hedges et al.[7] used a van Krevelen plot to constrain the probable chemical composition of marine algal biomass (Redfield biomass). Perdue and Ritchie[8] used a van Krevelen plot to display compositions of major classes of biomolecules and the chemical compositions of several hundred samples of fulvic acids, humic acids, and natural organic matter from fresh waters. The selected papers that have been cited are representative of hundreds of other papers in which van Krevelen plots have been used to understand organic matter in rocks, soils, sediments, and natural waters. The data used to construct those plots were, in probably all instances, bulk average chemical compositions of highly complex mixtures and not the chemical compositions of individual organic compounds.

With the advent of Fourier transform ion cyclotron resonance mass spectrometry (FTICR-MS), the highly complex mixtures of organic compounds that occur in nature finally began to yield some molecular-level compositional information. Specifically, molecular masses could be determined to such a high degree of accuracy that it was often possible to assign unique molecular formulae to individual peaks in a mass spectrum. Kim et al.[9] first used a van Krevelen plot to display and interpret the compositional data that are generated by FTICR-MS for a humic substance. An approximation of that van Krevelen plot has been generated by digitizing the data in Figure 3 of Kim et al.[9] and is given in Figure 1. This plot was profoundly different from the unstructured van Krevelen plots that are produced using bulk average compositional data. Kim et al.[9] recognized several linear trends in this plot. They attributed the nonparallel series of lines that converge at H/C =

Received: November 7, 2014
Accepted: February 4, 2015
Published: February 4, 2015

DOI: 10.1021/ac504165k
Anal. Chem. 2015, 87, 5079−5085

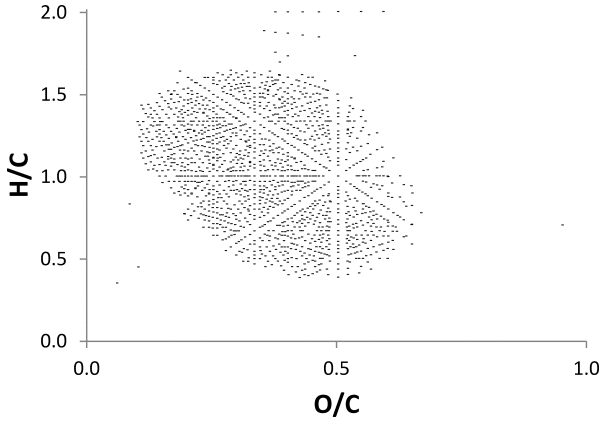


Figure 1. van Krevelen plot for McDonalds Branch dissolved organic matter.[9]

2 and O/C = 0 to processes by which molecular formulae gain/lose CH_2 groups. The line passing through the origin of the plot and through H/C = 1 and O/C = 0.5 was attributed to gain/loss of H_2O, vertical lines were attributed to gain/loss of H_2, and horizontal lines were attributed to gain/loss of O. The striking linear features of this van Krevelen plot inspired Hertkorn et al.[10] to make a similar plot for all possible combinations of C, H, and O within a molecular mass range of 200−700 Da. They found the same linear features throughout van Krevelen space, and they showed that a significant proportion of the theoretically possible combinations of C, H, and O can be found in Suwannee River fulvic acid.

van Krevelen plots clearly offer a powerful means of displaying and analyzing compositional data for highly complex mixtures, especially when molecular-level information is rendered from those mixtures by FTICR-MS. All prior studies in which van Krevelen plots have been utilized share the following characteristics: (1) only the positive quadrant of van Krevelen space, where H/C ≥ 0 and O/C ≥ 0, has been displayed and interpreted; (2) either bulk average chemical compositions have been plotted or, in the case of FTICR-MS data, molecular formulae covering a range of molecular masses have been superimposed in a van Krevelen plot.

The reader is now invited to explore simpler van Krevelen plots that contain only isobaric compounds of C, H, and O (compounds having the same nominal mass, NM) and to expand the view of van Krevelen space to include its negative quadrants, where H/C < 0 and/or O/C < 0. The value of this approach should soon become evident to the reader.

GENERATING MOLECULAR FORMULAE

The forthcoming figures in this paper utilize molecular formulae for compounds containing C, H, and O that were generated numerically using a simple algorithm. One such algorithm in the Pascal programming language and its output are given in Table 1.

This small program generates 53 molecular formulae that have a NM of 400 Da and which meet the additional constraints that 2 ≤ H ≤ 2C + 2 and 0 ≤ O ≤ C + 2. Numerous chemically invalid molecular formulae are also generated but are not displayed. The lower limit for H and the upper limit for O are somewhat arbitrary, but other limits are

Table 1. Pascal Program To Generate Isobaric Molecular Formulae and the Output List of Valid Molecular Formulae

```
PROGRAM Isobars;
{
Generates CHO molecular formulae in which
2 ≤ H ≤ 2C+2 and 0 ≤ O ≤ C+2.
}
VAR
    NM,C,H,O: WORD;

FUNCTION Valid: BOOLEAN;
BEGIN
    Valid:=(H>=2) AND (H<=2*C+2) AND (O>=0) AND (O<=C+2);
END;

BEGIN
    NM:=400;
    C:=NM DIV 12;
    REPEAT
        H:=NM-12*C;
        O:=0;
        IF Valid THEN WRITELN (C:6, H:6, O:6);
        WHILE (H>=16) AND (O<C+2) DO
        BEGIN
            H:=H-16;
            O:=O+1;
            IF Valid THEN WRITELN (C:6, H:6, O:6);
        END;
        C:=C-1;
    UNTIL NM>30*C+34; {NM > 12C + 1(2C+2) + 16(C+2)}
END.
```

OUTPUT FOR NM = 400 Da

C	H	O	C	H	O
33	4	0	23	44	5
32	16	0	23	28	6
31	28	0	23	12	7
31	12	1	22	40	6
30	40	0	22	24	7
30	24	1	22	8	8
30	8	2	21	36	7
29	52	0	21	20	8
29	36	1	21	4	9
29	20	2	20	32	8
29	4	3	20	16	9
28	48	1	19	28	9
28	32	2	19	12	10
28	16	3	18	24	10
27	44	2	18	8	11
27	28	3	17	36	10
27	12	4	17	20	11
26	40	3	17	4	12
26	24	4	16	32	11
26	8	5	16	16	12
25	52	3	15	28	12
25	36	4	15	12	13
25	20	5	14	24	13
25	4	6	14	8	14
24	48	4	13	20	14
24	32	5	13	4	15
24	16	6			

DOI: 10.1021/ac504165k
Anal. Chem. 2015, 87, 5079−5085

imposed by the rules of chemical bonding.[11] Molecular formulae for other NMs may be generated by modifying the assigned value of NM in the program. Chemically invalid molecular formulae may be displayed by changing the two instances of "IF Valid THEN" to "IF NOT Valid THEN".

VAN KREVELEN PLOTS FOR ISOBARIC MOLECULAR FORMULAE

The isobaric series of molecular formulae in Table 1 for NM = 400 Da are shown as filled circles in a van Krevelen plot in Figure 2. The open circles are a subset of the additional

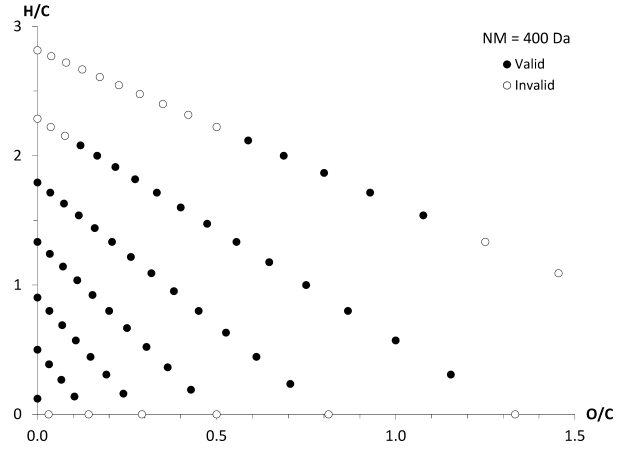


Figure 2. van Krevelen plot for isobaric CHO molecular formulae with a nominal mass of 400 Da. Filled and unfilled symbols represent chemically valid and invalid molecular formulae, respectively.

molecular formulae that are mathematically valid but do not meet the imposed chemical constraints that $2 \leq H \leq 2C + 2$ and $0 \leq O \leq C + 2$. They are included here only to aid the reader in visualizing linear patterns within the array of points. Qualitatively similar plots are obtained for other values of NM, with the density of compositions in van Krevelen space increasing with increasing NM.

This van Krevelen plot for a single isobaric series of molecular formulae reveals a surprisingly large number of linear patterns, some of which are more apparent than others. Starting with the array of points having the highest H/C for any given O/C and rotating clockwise, there are seven nonparallel lines that must intersect at one or more locations in the upper left negative quadrant of van Krevelen space. Likewise, starting with the array of points having the lowest H/C for any given O/C and rotating counterclockwise, there are 13 nonparallel lines that must intersect at one or more locations in the upper left or lower left negative quadrant of van Krevelen space. The reader may be able to identify several additional linear patterns within this quite remarkable figure. Similar trends for NM = 576 Da were identified by Hertkorn et al.[10] and published as Figure 3A in the Supporting Information for that paper.

It may be surprising to note that these lines in Figure 2 are not evident when multiple isobaric series are combined in a single van Krevelen plot. Likewise, none of the lines in Figure 1 that were observed by Kim et al.[9] and attributed to gain/loss of CH_2, H_2O, H_2, and O are observed in Figure 2. This is appropriate, because gain/loss of CH_2, H_2O, H_2, and O must cause the NM to change, and all compositions in Figure 2 have the same NM.

EXPLORING THE NEGATIVE QUADRANTS OF VAN KREVELEN SPACE

The chemical compositions of molecules necessarily plot in the positive quadrant of van Krevelen space (O/C ≥ 0, H/C ≥ 0)—the only portion of that space which has heretofore been explored. The aforementioned series of nonparallel lines passing through those compositions appear to converge in the negative quadrants of van Krevelen space. Three such sets of linear patterns are highlighted in Figure 3, where only valid

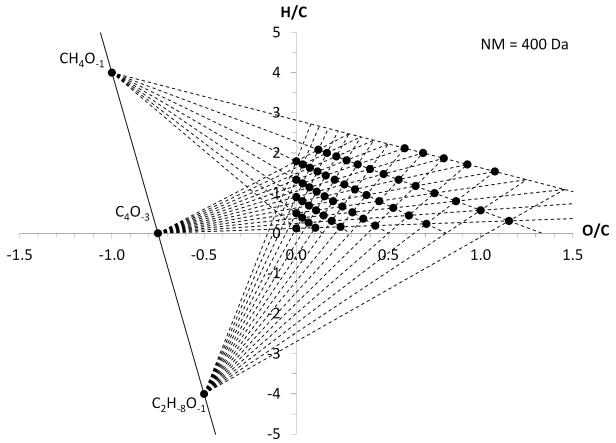


Figure 3. van Krevelen plot for isobaric CHO molecular formulae with a nominal mass of 400 Da, including the negative quadrants of van Krevelen space.

molecular formulae are plotted and where all four quadrants of van Krevelen space are viewed. The first set of linear patterns that were identified in the previous paragraph all intersect at a single point in the upper left quadrant of van Krevelen space at H/C = 4 and O/C = −1. The moiety responsible for this point in van Krevelen space will be represented by a molecular formula of CH_4O_{-1}. The second set of linear patterns that were identified in the previous paragraph all intersect at a single point in van Krevelen space at H/C = 0 and O/C = −3/4. The moiety having this composition will be represented by a molecular formula of C_4O_{-3}. Finally, a third set of previously unmentioned linear patterns all intersect at a single point in the lower left quadrant of van Krevelen space at H/C = −4 and O/C = −1/2. The moiety having this composition will be represented by a molecular formula of $C_2H_{-8}O_{-1}$.

Once the view of van Krevelen plots is expanded in Figure 3 to include the negative quadrants of that compositional space, it seems natural and appropriate to use molecular formulae to represent the moieties that are found therein. The moieties are well-known and are often invoked to explain patterns in the FTICR mass spectra of complex mixtures such as humic acids, fulvic acids, and natural organic matter. In that literature, however, they are not represented by molecular formulae but instead by equivalent phrases. Examples include "$+CH_4/-O$",[10] "CH_4 vs O",[12] "replacement of CH_4 by O",[13] etc. The paper by Hertkorn et al.,[10] in particular, specifically addresses the use of these moieties (and others) in exploration and interpretation of the compositional space of natural organic matter.

All three of the moieties whose compositions plot in the negative quadrants of van Krevelen space (Figure 3) have NM = 0; however, the exact masses (EMs) of CH_4O_{-1}, C_4O_{-3}, and $C_2H_{-8}O_{-1}$ are +36.3855, +15.2561, and −57.5149 mDa, respectively. These and other related moieties will be called

DOI: 10.1021/ac504165k
Anal. Chem. 2015, 87, 5079−5085

low-mass moieties (or LMMs), because their EM values are small but nonzero. The three LMMs have compositions that fall on a straight line that never enters the positive quadrant of van Krevelen space. The equation of that line (H/C = −12 − 16(O/C)) is derived easily from the mass balance equation for a moiety with a NM of zero (12C + 1H + 16O = 0). One additional moiety that should be mentioned is found at the intercept of the moiety line, i.e., at H/C = −12 and O/C = 0. The moiety having this composition will be represented by a molecular formula of CH_{-12}. Yet another series of nonparallel lines pass through the compositions in Figure 3 and intersect at this moiety, but those lines and their intersection point have been omitted to simplify Figure 3.

Because a straight line is defined by two points, it is possible to use the molecular formulae and EMs of any two LMMs to derive the corresponding properties of other LMMs. For example, the third moiety ($C_2H_{-8}O_{-1}$) in Figure 3 may be considered formally as a linear combination of CH_4O_{-1} and C_4O_{-3}:

$$-2CH_4O_{-1} + C_4O_{-3} = C_2H_{-8}O_{-1} \quad (1)$$

Its EM of −57.5149 mDa is simply −2(+36.3855) +1-(+15.2561). The $C_{43}H_{-20}O_{-31}$ moiety, which has an EM of +1.1474 mDa, was introduced by Hertkorn et al.[10] That moiety may be constructed formally by

$$-5CH_4O_{-1} + 12C_4O_{-3} = C_{43}H_{-20}O_{-31} \quad (2)$$

Its EM of +1.1474 mDa is simply −5(+36.3855) +12-(+15.2561). Tables of low-mass moieties and their exact masses[10] can be simplified or eliminated, because all moieties containing only C, H, and O are linear combinations of two LMMs such as CH_4O_{-1} and C_4O_{-3}.

USING LOW-MASS MOIETIES TO EXPLORE VAN KREVELEN SPACE

Figure 4 is a simplified version of Figure 3, in which only two LMMs are shown and only the upper quadrants of van Krevelen space are viewed. Figure 4 will be used to examine more thoroughly the use of LMMs for exploration and interpretation of van Krevelen space. Compositions of molecular formulae in the positive quadrant of van Krevelen space that fall on a straight line passing through CH_4O_{-1} in the upper left negative quadrant differ compositionally from one another by the composition of CH_4O_{-1}. This relationship can be illustrated using the molecular formula $C_{13}H_{20}O_{14}$, whose location is at the upper right end of the heavy black line in Figure 4. The line connecting $C_{13}H_{20}O_{14}$ and the CH_4O_{-1} moiety can be considered as a two-component mixing line and will be designated as a CH_4O_{-1} mixing line. If one mole of the CH_4O_{-1} moiety is added to (or subtracted from) $C_{13}H_{20}O_{14}$, the resulting molecular formula falls on the mixing line in the direction toward (or away from) CH_4O_{-1}. A range of molecular formulae such as $C_{12}H_{16}O_{15}$, $C_{13}H_{20}O_{14}$, $C_{14}H_{24}O_{13}$, $C_{15}H_{28}O_{12}$, $C_{16}H_{32}O_{11}$, $C_{17}H_{36}O_{10}$, and $C_{18}H_{40}O_9$ can be obtained in this manner. Only five of these molecular formulae are in Figure 4. The first formula in this list is not found because O > C + 2, and the last formula is not found because H > 2C + 2. The EM of $C_{13}H_{20}O_{14}$ is 400.0853 Da, and the EM of CH_4O_{-1} is +36.3855 mDa. Adjacent points along a CH_4O_{-1} mixing line differ in mass by the EM of CH_4O_{-1}, with EM increasing in the direction of CH_4O_{-1}. *It is very important to understand that any two valid molecular formulae on a CH_4O_{-1} mixing line must differ in EM by an integer multiple of the mass of CH_4O_{-1} (+36.3855 mDa).*

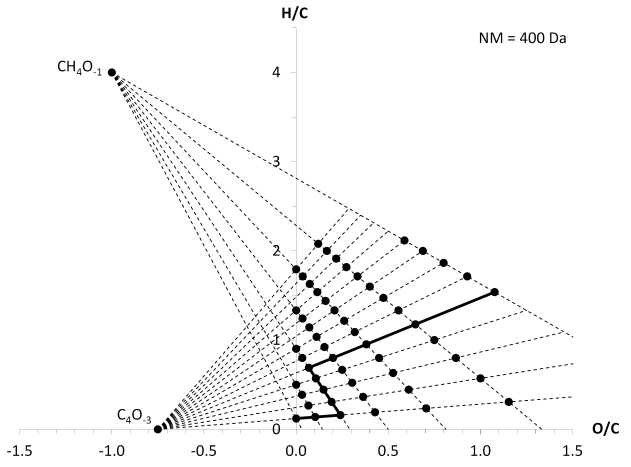


Figure 4. van Krevelen plot for isobaric CHO molecular formulae with a nominal mass of 400 Da, including the top left negative quadrant of van Krevelen space.

In a similar fashion, the line connecting $C_{13}H_{20}O_{14}$ and the C_4O_{-3} moiety can be considered as a two-component mixing line and will be designated as a C_4O_{-3} mixing line. If the C_4O_{-3} moiety is added to (or subtracted from) $C_{13}H_{20}O_{14}$, the resulting molecular formula falls on the next CH_4O_{-1} mixing line in the direction toward (or away from) C_4O_{-3}. In this particular case, it is not possible to remove C_4O_{-3} because the resulting molecular formula ($C_9H_{20}O_{17}$) would be chemically invalid (O > C + 2). The EM of $C_{13}H_{20}O_{14}$ is 400.0853 Da, and the EM of C_4O_{-3} is +15.2561 mDa. Adjacent points along a C_4O_{-3} mixing line differ in mass by the EM of C_4O_{-3}, with EM increasing in the direction of C_4O_{-3}. Any two valid molecular formulae on a line leading to C_4O_{-3} must differ in mass by an integer multiple of the mass of C_4O_{-3} (+15.2561 mDa).

When working with isobaric series of molecular formulae, the CH_4O_{-1} and C_4O_{-3} moieties can be used together to navigate to any valid molecular formula in van Krevelen space. For example, $C_{13}H_{20}O_{14}$ can be converted into the isobaric hydrocarbon having the lowest H/C ratio by stepping through the grid that is defined by the two LMMs. Starting with the known molecular formula of $C_{13}H_{20}O_{14}$ (EM = 400.0853 Da), any path along the grid by which six C_4O_{-3} moieties are added and four CH_4O_{-1} moieties are subtracted will lead to the hydrocarbon. One possible path is the heavy black line in Figure 4. The formula and EM of the hydrocarbon are easily generated:

$$C_{13}H_{20}O_{14} - 4CH_4O_{-1} + 6C_4O_{-3} = C_{33}H_4 \quad (3)$$

The EM of the hydrocarbon (400.0313 Da) is simply 400.0853 − 4(+0.0363855) + 6(+0.0152561). (Note that the EMs of the LMMs were converted here from millidaltons to daltons.) In this fashion, the molecular formulae and EM values of all isobaric compounds of C, H, and O can be generated quickly and systematically from the molecular formulae and EMs of one member of that series and the two LMMs (CH_4O_{-1} and C_4O_{-3}).

USING LOW-MASS MOIETIES TO ASSIGN MOLECULAR FORMULAE

Once the use of LMMs to navigate van Krevelen space is understood, it becomes possible to realize the full power of

DOI: 10.1021/ac504165k
Anal. Chem. 2015, 87, 5079−5085

LMMs to facilitate the interpretation of FTICR mass spectra. Suppose an EM has been determined experimentally. What is the molecular formula that corresponds to this EM?

General Procedure for Determining a Molecular Formula from an Exact Mass. The procedure will be described here in detail and is illustrated in Table 2. (1) Convert the EM to the NM of an isobaric series. (2) Calculate the molecular formula and EM of the hydrocarbon in this series that has the maximum number of moles of C in its formula. This will be the initial reference molecular formula in the search for the correct molecular formula. (3) Check to see if the unknown molecular formula lies on the CH_4O_{-1} mixing line on which the reference molecular formula is found. All molecular formulae on this mixing line differ in EM by an integer multiple of the EM of CH_4O_{-1}. If this is the case, then add/subtract the required number of moles of CH_4O_{-1} to/from the reference molecular formula to reach the correct EM and molecular formula. Otherwise, subtract one mole of C_4O_{-3} from the reference molecular formula to generate a new reference molecular formula on the next CH_4O_{-1} mixing line. (4) Repeat step 3 until the correct molecular formula is found or until all valid compositional space has been explored.

Table 2. Using Low-Mass Moieties To Determine the Molecular Formula That Has an Exact Mass of 400.1006 Da and a Nominal Mass of 400 Da

step of procedure	NM, Da	C	H	O	EM, Da	no. of CH_4O_{-1} moieties
find hydrocarbon with maximum C	400	33	4	0	400.0313	1.90
move to the next line by subtracting C_4O_{-3}	400	29	4	3	400.0160	2.32
move to the next line by subtracting C_4O_{-3}	400	25	4	6	400.0008	2.74
move to the next line by subtracting C_4O_{-3}	400	21	4	9	399.9855	3.16
move to the next line by subtracting C_4O_{-3}	400	17	4	12	399.9703	3.58
move to the next line by subtracting C_4O_{-3}	400	13	4	15	399.9550	4.00
get the final formula by adding $4CH_4O_{-1}$	400	17	20	11	400.1006	0.00

Specific Example of Assigning a Molecular Formula to an Exact Mass. A specific example is presented in Table 2, where the above four-step procedure is followed to determine the molecular formula of a molecule having an EM of 400.1006 Da. The problem is solved in a total of seven simple steps. The first step requires the NM to be divided by 12 to obtain the maximum whole number of C, with H used for all remaining mass. Then the difference in mass between the correct molecular formula (400.1006 Da) and the hydrocarbon (400.0313 Da) is divided by the EM of CH_4O_{-1} (0.0363855 Da) to obtain a ratio of 1.90, which is not an integer. The correct molecular formula cannot exist on this CH_4O_{-1} mixing line. The next five steps involve subtraction of the C_4O_{-3} moiety from the reference molecular formula to generate another reference molecular formula on the next CH_4O_{-1} mixing line. At that point, the mass difference between the correct molecular formula (400.1006 Da) and the reference molecular formula (399.9550 Da) equals 4.00 CH_4O_{-1} moieties. The correct molecular formula ($C_{17}H_{20}O_{11}$) is obtained by adding those four CH_4O_{-1} moieties to the reference formula. The steps in Table 2 may be carried out without reference to a van Krevelen plot. Even so, the reader might wish to start at the composition of the hydrocarbon in the lower left corner of Figure 4 and follow the steps in Table 2 to locate the correct molecular formula. Low-mass moieties make this task trivially easy, so much so that students have been able to solve this problem easily, accurately, and quickly with a hand-held calculator during an in-class examination.

Computational Efficiency of Using LMMs To Assign Molecular Formulae. In the "brute force" approach of assigning molecular formulae to EM values, all possible values of C, H, and O are tried systematically until the correct molecular formula is found. With reference to Table 1, the maximum numbers of moles of C, H, and O in the molecular formulae of the 400 Da isobaric series are 33, 52, and 15, respectively. The following snippet of Pascal code illustrates a brute-force approach that might be taken:

```
FOR W:=1 TO 33 DO {C loop}
BEGIN
   FOR X:=2 TO 52 DO {H loop in steps of 2}
   BEGIN
      FOR Y:=0 TO 15 DO {O loop}
      BEGIN
         {Compare the EM of the target molecule
          with the EM of CWHXOY.}
      END;
   END;
END;
```

Note that H must be an even number in molecular formulae containing only C, H, and O. This inefficient approach might require 33 × 26 × 16 (13728) loop steps, in addition to the calculations that must be done in the innermost loop. In this particular case, the correct molecular formula ($C_{17}H_{20}O_{11}$) would be found on the 2040th loop. The size of the computational problem can be reduced by eliminating the C loop (the largest) and using mass balance to obtain the moles of C, thus reducing the maximum number of loops to 26 × 16 (416). In this case, the correct molecular formula ($C_{17}H_{20}O_{11}$) would be found on the 120th loop.

The moiety-based approach required only seven steps (see Table 2), yielding the correct molecular formula as much as 291 (=2040/7) times more efficiently than the three-loop brute-force approach and 17 (=120/7) times more efficiently than the two-loop brute-force approach. These increases in the efficiency of fitting molecular formulae to EM data should be reflected in faster execution times, which have, in fact, been realized and are the topic of a companion paper by the same authors.[14] That paper expands the moiety-based algorithm that is described here to include ^{13}C, ^{14}N, ^{31}P, and ^{32}S. The expanded moiety-based algorithm provides a very fast deterministic method of assigning all possible valid molecular formulae to an EM, and it can be complemented by several heuristic algorithms that aim to select the correct molecular formula from among the possibilities.

FAMILY SCORES, LOW-MASS MOIETIES, AND VAN KREVELEN PLOTS

Stenson et al.[13] combined the Z^* sorting parameter of Hsu et al.,[15] unsaturation (double bond equivalents), and oxygen content of molecular formulae and discovered that the vast majority of molecular formulae with even NMs in Suwannee River fulvic acid could be described by two simple equations:

DOI: 10.1021/ac504165k
Anal. Chem. 2015, 87, 5079−5085

$$\mathrm{DBE} - \mathrm{O} = -0.5Z^* + 1 \quad (4)$$

$$\mathrm{DBE} - \mathrm{O} = -0.5Z^* - 6 \quad (5)$$

where DBE is double bond equivalents or unsaturation. Other similar equations were found for molecular formulae having odd NMs because they contained an odd number of ^{13}C and/or ^{14}N. From a careful reading of this paper, the discovery seems to have been serendipitous and the equations seem to be empirical. In subsequent papers, Stenson[16,17] has rearranged eqs 4 and 5 and combined them into a single equation, introducing the term "family score":

$$\text{family score} = {}^1\!/_2 Z^* + (\mathrm{DBE} - \mathrm{O}) \quad (6)$$

In these three papers, Stenson and co-workers have found molecular formulae having family scores of +8, +1, −6, −13, and −20. These intriguing results are connected directly to the low-mass moieties that are the focus of this paper. To describe these connections, it is helpful to begin with a fundamental derivation of family scores for molecular formulae containing C, H, and O and then to re-examine Figure 4 in terms of family scores and LMMs.

Derivation of Family Scores for Molecular Formulae Containing C, H, and O. In an isobaric series of molecular formulae containing C, H, and O, the NM is given by NM = 12C + 1H + 16O and DBE is given by DBE = $^1/_2$(2C + 2 − H). Solving the DBE equation for H and substituting into the equation for NM yields

$$\mathrm{NM} = 12\mathrm{C} + (2\mathrm{C} + 2 - 2\mathrm{DBE}) + 16\mathrm{O} \quad (7)$$

Now a minor redistribution of terms is used to prepare for calculation of Z^*:

$$\mathrm{NM} = 14\mathrm{C} + 14\mathrm{O} + [2 - 2\mathrm{DBE} + 2\mathrm{O}] \quad (8)$$

$$Z^* = \mathrm{modulus}(\mathrm{NM}/14) - 14 \quad (9)$$

Substituting from eq 8 for NM and noting that 14C + 14O is evenly divisible by 14 and thus does not contribute to the modulus yields

$$Z^* = \mathrm{modulus}([2 - 2\mathrm{DBE} + 2\mathrm{O}]/14) - 14 \quad (10)$$

Now another nonintuitive step is used to process the modulus expression. The modulus is simply the remainder following integer division and can be expressed as [2 − 2DBE + 2O] − 14Q, where Q is the quotient of integer division in the modulus expression (Q = 0, ± 1, ± 2, ± 3, etc.). Thus

$$Z^* = [2 - 2\mathrm{DBE} + 2\mathrm{O}] - 14Q - 14 \quad (11)$$

Rearranging eq 11 and solving for Q

$$Q = -([{}^1\!/_2 Z^* + (\mathrm{DBE} - \mathrm{O})] + 6)/7 \quad (12)$$

The term in brackets is the family score from eq 6.[17] The family scores that have been observed by Stenson and co-workers (+8, +1, −6, −13, and −20) correspond to Q values of −2, −1, 0, +1, and +2, respectively. Although the family score of a particular molecular formula must be calculated from eq 6,[17] the allowed values of the family score may be calculated for any value of Q using a rearranged version of eq 12:

$$\text{family score} = -(6 + 7Q) \quad (13)$$

Derivation of Family Scores for More Complex Molecular Formulae. The approach described in eqs 7−12 is easily extended to the more complex case of molecular formulae that also contain ^{13}C, ^{14}N, ^{31}P, and ^{32}S (and any other elements of interest). Using E to represent ^{13}C, more comprehensive equations are used for NM and DBE:

$$\mathrm{NM} = 12\mathrm{C} + 1\mathrm{H} + 16\mathrm{O} + 13\mathrm{E} + 14\mathrm{N} + 32\mathrm{S} + 31\mathrm{P} \quad (14)$$

$$\mathrm{DBE} = {}^1\!/_2[2(\mathrm{C} + \mathrm{E}) + 2 + \mathrm{N} + \mathrm{P} - \mathrm{H}] \quad (15)$$

From this point on, the derivation follows eqs 7−12. Equation 10 becomes

$$Z^* = \mathrm{modulus}([2 - 2\mathrm{DBE} + 2\mathrm{O} + 4\mathrm{S} + 4\mathrm{P} + \mathrm{N} + \mathrm{E}]/14) - 14 \quad (16)$$

The derivation proceeds as before to finally obtain

$$Q = -([{}^1\!/_2 Z^* + (\mathrm{DBE} - \mathrm{O})] - [2\mathrm{S} + 2\mathrm{P} + {}^1\!/_2\mathrm{N} + {}^1\!/_2\mathrm{E}] + 6)/7 \quad (17)$$

The original family score from Stenson and co-workers[12,16,17] is the leftmost bracketed term in eq 17. If the simple relationship between family score and Q in eq 13 is to be maintained, then a broader definition of family score is required for more complex molecular formulae. We recommend a redefinition of family score as

$$\text{family score} = [{}^1\!/_2 Z^* + (\mathrm{DBE} - \mathrm{O})] - [2\mathrm{S} + 2\mathrm{P} + {}^1\!/_2\mathrm{N} + {}^1\!/_2\mathrm{E}] \quad (18)$$

which is then consistent with eq 13 and unifies all molecular formulae into a single, easily predicted range of family scores.

Family Scores, Low-Mass Moieties, and van Krevelen Plots. Family scores have been calculated using eq 6 for each of the molecular formulae in Table 1. Figure 5 is constructed from Figure 4 simply by labeling each molecular formula with its family score. It becomes clear that all molecular formulae that lie on the same CH_4O_{-1} mixing line have the same family score. Starting with the line having the highest H/C at any given O/C and rotating clockwise through the remaining lines, the values of Q that correspond to the family scores are +1, 0, −1, −2, −3, −4, and −5. Although we prefer to use the simpler Q series, the

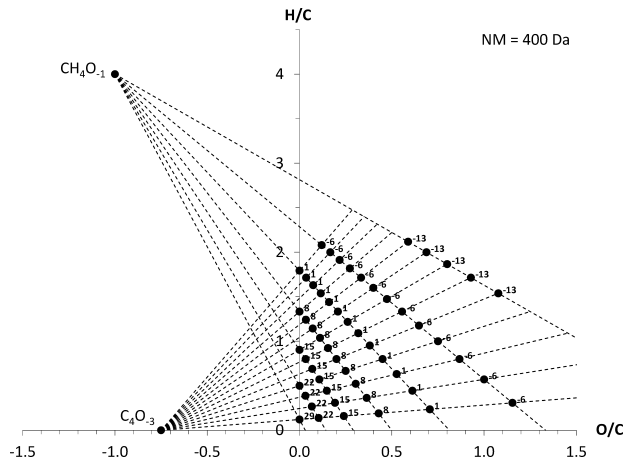


Figure 5. van Krevelen plot for isobaric CHO molecular formulae with a nominal mass of 400 Da, including the top left negative quadrant of van Krevelen space and showing the family score for each molecular formula.

DOI: 10.1021/ac504165k
Anal. Chem. 2015, 87, 5079−5085

molecular formulae in Figure 5 are labeled with family scores to facilitate comparison with the work of Stenson and colleagues. As noted earlier, Stenson et al.[13] and Stenson[16,17] observed family scores of −20, −13, −6, +1, and +8, respectively. They did not observe the family scores of +15, +22, and +29 that represent molecular formulae that contain very low O and very high DBE (the lower left corner of the van Krevelen plot). The isobaric series for NM = 400 Da does not include any valid compositions for which the family score is −20.

Stenson and co-workers discovered the relationships that led to the concept of family scores while working with molecular formulae for real FTICR-MS data sets containing relatively large ranges of NM. The analysis in this paper has been restricted to a single isobaric series of molecular formulae because the patterns that are revealed clearly in Figure 5 would be obscured if multiple isobaric series were superimposed in a single van Krevelen plot, just as is the case when entire FTICR-MS data sets are superimposed in a single van Krevelen plot. Nonetheless, the patterns revealed in this paper and the relationship between LMMs and family scores are observable for any isobaric series. The reader can easily confirm this assertion by generating molecular formulae containing C, H, and O for an isobaric series (use the computer program in Table 1), plotting those compositions in a van Krevelen plot, and calculating family scores.

With reference to LMMs, all molecular formulae that fall on the same CH_4O_{-1} mixing line have the same family score and Q value, and the individual members of that series differ by gain/loss of the CH_4O_{-1} moiety. Both family scores and Q values are changed by gain/loss of the C_4O_{-3} moiety. For example, if a molecular formula with a family score of −13 (Q = +1) gains C_4O_{-3}, the resulting molecular formula will lie on the line with a family score of −6 (Q = 0). To summarize these findings for an isobaric series of molecular formulae, the CH_4O_{-1} moiety interconverts molecular formulae that have the same family score, and the C_4O_{-3} moiety interconverts molecular formulae having different family scores.

CONCLUSIONS

A van Krevelen plot for a single isobaric series of molecular formulae reveals numerous linear trends that are not observable in a conventional van Krevelen plot for a complex mixture. When the view of van Krevelen space is expanded to include its negative quadrants, these trend lines are found to converge at coordinates that correspond to moieties having zero nominal mass and very small exact mass. Five such low-mass moieties have been mentioned specifically in this paper (CH_4O_{-1}, C_4O_{-3}, $C_2H_{-8}O_{-1}$, CH_{-12}, and $C_{43}H_{-20}O_{-31}$). All such moieties are linear combinations of the compositions of any two moieties and have compositions that plot on the line H/C = −12 − 16(O/C), which lies entirely in the negative quadrants of van Krevelen space. Two of the LMMs (CH_4O_{-1} and C_4O_{-3}) have been used to explain compositional relationships within an isobaric series, to interconvert molecular formulae within an isobaric series, and to assign a unique molecular formula to a given exact mass. The moiety-based approach to assigning molecular formulae is far more efficient and faster than brute-force searches over the allowed ranges of C, H, and O. The authors have coupled the moiety-based approach for C, H, and O with conventional nested loops for ^{13}C, N, S, and P in a companion paper.[14]

It has been shown that all molecular formulae on a CH_4O_{-1} mixing line have the same family score and that all molecular formulae on a C_4O_{-3} mixing line have different family scores. Finally, the definition of family score has been extended in a consistent manner to include molecular formulae that contain heteroatoms and ^{13}C.

ASSOCIATED CONTENT

Supporting Information

van Krevelen plot of CHO molecular formulae between 150 and 1000 Da, with an example for visualizing isobaric series (XLS). This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: emperdue@bsu.edu. Phone: +0017652858096.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) van Krevelen, D. W. *Fuel* **1950**, *29*, 269−284.
(2) Durand, B.; Monin, J. C. In *Kerogen: Insoluble Organic Matter from Sedimentary Rocks*; Durand, B., Ed.; Editions Technip: Paris, 1980; pp 113−142.
(3) Orlov, D. S. *Humus Acids of Soils, Russian Translation Series 35*; A. A. Balkema: Rotterdam, The Netherlands, 1985; pp 35−70.
(4) Reuter, J. H.; Perdue, E. M. In *Mitteilungen aus dem Geologisch-Paläontologischen Institut der Universität Hamburg, Heft 56*; Degens E. T., Krumbein W. E., Prashnowsky A. A., Eds.; Das Institut: Hamburg, 1984; pp 249−262.
(5) Visser, S. A. *Environ. Sci. Technol.* **1983**, *17*, 412−417.
(6) Sun, L.; Perdue, E. M.; Meyer, J. L.; Weis, J. *Limnol. Oceanogr.* **1997**, *42*, 714−721.
(7) Hedges, J. I.; Baldock, J. A.; Gélinas, Y.; Lee, C.; Peterson, M. L.; Wakeham, S. G. *Mar. Chem.* **2002**, *78*, 47−63.
(8) Perdue, E. M.; Ritchie, J. R. In *Surface and Ground Water, Weathering, and Soils, Vol. 5, Treatise on Geochemistry*; Drever, J. I., Holland, H. D., Turekian, K. K., Eds.; Elsevier-Pergamon: Oxford, U.K., 2003; pp 273−318.
(9) Kim, S.; Kramer, R. W.; Hatcher, P. G. *Anal. Chem.* **2003**, *75*, 5336−5344.
(10) Hertkorn, N.; Frommberger, M.; Witt, M.; Koch, B.; Schmitt-Kopplin, Ph.; Perdue, E. M. *Anal. Chem.* **2008**, *80*, 8908−8919.
(11) Senior, J. K. *Am. J. Math.* **1951**, *73*, 663−689.
(12) Stenson, A. C.; Landing, W. M.; Marshall, A. G.; Cooper, W. T. *Anal. Chem.* **2002**, *74*, 4397−4409.
(13) Stenson, A. C.; Marshall, A. G.; Cooper, W. T. *Anal. Chem.* **2003**, *75*, 1275−1284.
(14) Green, N. W.; Perdue, E. M. *Anal. Chem.* **2015**, DOI: 10.1021/ac504166t.
(15) Hsu, C. S.; Qian, K.; Chen, Y. C. *Anal. Chim. Acta* **1992**, *264*, 79−89.
(16) Stenson, A. C. *Environ. Sci. Technol.* **2008**, *42*, 2060−2065.
(17) Stenson, A. C. *Rapid Commun. Mass Spectrom.* **2009**, *23*, 465−476.

DOI: 10.1021/ac504165k
Anal. Chem. 2015, 87, 5079−5085