Online Computation of Molecular Formulas from Mass Number

Joshua Lederberg
Department of Genetics
Stanford University School of Medicine
Stanford, California

The calculation of hypotheses of molecular formulas to match a given mass number is a fussy and error-prone chore, notwithstanding its steps are the simplest arithmetic. A number of tables have been published to alleviate this burden for mass spectrometrists (1, 2, 3, 4). Better still, computer programs have been written to read input punched cards and return printed lists of the corresponding fo~mulas (2, 5, 6).

The advent of time-shared computer facilities portends that chemists will take increasing advantage of such programs as workaday tools. The expected intensity of application stresses the desirability of an efficient algorithm. At the same time this should take a minimum of instructions so that it can be accessed and executed in a single pass in a time-sharing environment.

A program to meet these specifications has been written for the time-sharing system on the Q-32 computer at System Development Corporation, Santa Monica, California. It was composed and debugged and is routinely called and executed over a 400 mile teletype link from Stanford. Including dialup and logging in, the typical turnaround time is one minute; only a few milliseconds are taken up by the actual computation.

The program is written as a single block in the JOVIAL language, a dialect of ALGOL, readily translatable into PL-I or FORTRAN. It can be punched on 30 cards. Copies will be gladly furnished by the author on request.

The algorithm includes a number of shortcuts to eliminate fruitless searches. Mass numbers are rescaled to the base, $H^1 = 1.0000$. Hence the rescaled weights will differ from a nominal integer by a relative mass defect which varies for each atom of the other elements (C = 12 - 0.09317, N = 14 - 0.10565, O = 16 - 0.12927), but is zero for H. The scan over H values can then be limited by calculations on the mass defect, and by a valence rule (allowing for one excess proton) $H \leq 3 + 2C + N$ (7). H must also have the same parity as the nominal mass number.

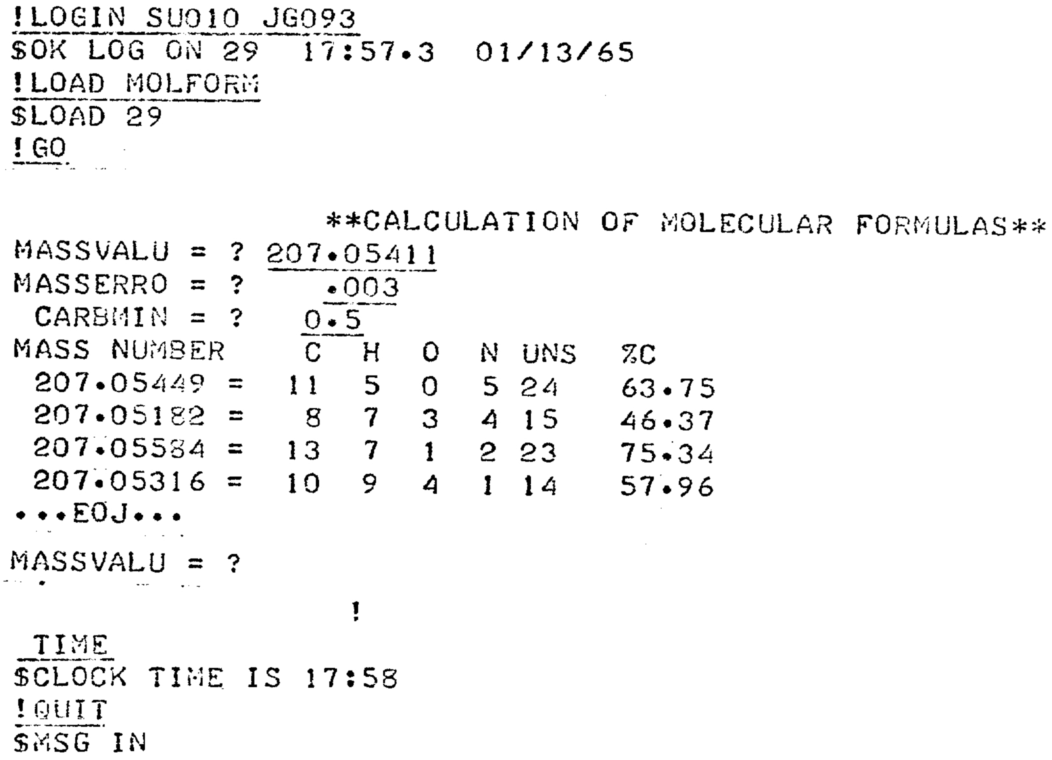
Within an H-scan loop (FOR H = (HMIN, 2, HMAX)); a scan for C is set up that is also limited by the results of a similar analysis. It is also constrained in advance by a CMIN read from the teletype (to limit excessive output of irrelevant forms) and by the requirement that the remaining mass be allocatable to the sum RWT = 14(N) + 16(O). This lends itself to some integer arithmetic by the transformation RWT = 14(N + O) + 2(O); hence RWT/14 - (N + O) with a remainder of 2(O), as the first solution. So long as the condition $(N + O \geq (O)$ can be fulfilled, i.e., $N \geq$ zero, the correspondence $8(N) \equiv 7(O)$ leads to additional trial solutions. Each of these is then tested for agreement with the given mass value ± mass error bounds, and if it passes the result is printed.

At least on some machines time could be saved by the elimination of floating-point arithmetic and by some additional use of look-up tables, but the present form of the program probably represents a reasonable balance of programming effort vs. efficiency. The algorithm is already rather more efficient than those previously reported. On the Q-32 the ad hoc revision of the symbolic program for special purposes is quite easy. For production tests it is stored as a precompiled binary text of 1704 words.

A self-explanatory printout of a complete test run is appended.

There is of course no theoretical obstacle to wiring the mass spectrometer directly to the computer; a closed loop operation is the aim of current work in this laboratory.

Figure 1. A sample run on data for dihydroxanthurenic acid (8). Other evidence for the presence of -C = O, -COOH, -OH, and N narrows the choice to $C_{10}H_9NO_4$. The text typed by the user has been underlined. The remaining text is the Q-32's response.



```
!LOGIN SU010 JG093
$OK LOG ON 29  17:57.3  01/13/65
!LOAD MOLFORM
$LOAD 29
!GO

                **CALCULATION OF MOLECULAR FORMULAS**
MASSVALU = ? 207.05411
MASSERRO = ?    .003
 CARBMIN = ?   0.5
MASS NUMBER    C  H  O  N UNS  %C
 207.05449 =  11  5  0  5 24   63.75
 207.05182 =   8  7  3  4 15   46.37
 207.05584 =  13  7  1  2 23   75.34
 207.05316 =  10  9  4  1 14   57.96
...EOJ...
MASSVALU = ?
                 !
 TIME
$CLOCK TIME IS 17:58
!QUIT
$MSG IN
```

Figure 2. Peptidolipin-NA (9). This datum was run at the suggestion of Professor C. Djerassi, to illustrate the scope of this technique. Any printed tables large enough to include such large molecules would be prohibitively voluminous. Even here, to limit the output, the program was revised slightly (from the teletype keyborad: this took about five minutes to implement) as indicated by the headings.

The last three items on the output list are perhaps equally plausible in the context of a substituted oligopeptide. Further analytical information (10) N=10.2%, gives a clear decision in favor of $C_{50}H_{89}O_{11}N_7$.

```
                         **MOLECULAR FORMULAS::PARENT MOLECULAR IONS
MASSVALU = ? 963.6599
MASSERRO = ?     .0030
 CARBMIN = ? 20
  NMINIM = ?  7
  OMINIM = ?  7
MASS NUMBER     C  H  O  N UNS   %N
 963.65714 =  27 77  8 31 10   45.05
 963.65982 =  31 81 10 25  8   36.33
 963.66251 =  35 85 12 19  6   27.61
 963.65932 =  46 85  9 13 22   18.89
 963.65799 =  45 89 13  9 12   13.08
 963.66201 =  50 89 11  7 20   10.17
...EOJ...
 MASSVALU = ?


                          !QUIT



  $MSG IN.
```

These studies are supported by contract No. AF 49(638)-1599 from the Air Force and grant No. NsG 81-60 from the National Aeronautics and Space Administration.

LITERATURE CITED

(1) Beynon, J. H., Williams, A. E., "Mass and Abundance Tables for Use in Mass Spectrometry", Elsevier, Amsterdam, 1963.

(2) Lederberg, J., "Computation of Molecular Formulas for Mass Spectrometry", Holden-Day, 1964.

(3) Tunnicliff, D. D., and Wadsworth, P. A., and Schissler, D. O. Anal. Chem. 37, 543 (1965).

(4) Van Katwijk, J., Appl. Spectroscopy 18, 102 (1964).

(5) Usher, D. A., Gougoutas, J. Z., and Woodward, R. B, Anal. Chem. 37, 330 (1965).

(6) Lederberg, J., and Wightman, M. A subalgol program for calculation of molecular compositional formulas from mass spectral data. NASA, Sci. and Tech. Aerospace Reports (STAR) No. N66-11689. (1964)

(7) Lederberg, J., and Wightman, M. Anal. Chem 36, 2365. (1964)

(8) Brown, K. S., J. Am. Chem. Soc. 87, 4202. (1965).

(9) Barber, M., Wolstenholme, W. A. and Lederer, E. Tetrahedron Letters 18, 1331. (1965)

(10) Guinand, M. Michel, G., and Lederer, E. C. R. Acad. Sci., 259, 1267 (1964)

BALGOL

updated ALGOL version to follow

```
        400  000000HWANG                 COMPARE WITH 360/5
                 STANFORD UNIVERSITY COMPILER -- VERSION OF  9/13/66

INS AT 163712

  144... COMMENT BALGOL PROGRAM TO CALCULATE MOLECULAR FORMULAS FROM A HIGH RESO
  144... LUTION MASS VALUE. THIS VERSION READS THE MASSVALUE, ITS ERROR, AND THE
  144... MINIMUM VALUE FOR CARBON CONTENT. THIS CAN BE STATED EITHER AS AN IN-
  144... TEGER NUMBER OF CARBON ATOMS OR AS A FRACTIONAL PROPORTION OF CARBON BY
  144... WEIGHT. EXCEPT FOR MINOR SYNTACTICAL DIFFERENCES, THIS IS THE SAME AS
  144... THE CORRESPONDING PROGRAM IN TINT/JOVIAL, MOLCALC.
  144... THE MAJOR ADAPTATIONS FOR OTHER LANGUAGAGES ARE PROBABLY  THE READ AND
  144... WRITE. $
  144...  FLOATING OTHERWISE$
  144... COMMENT TRUNCATION BY INTEGER ARITHMETIC IS INTENDED AND IMPORTANT$
  144...    COMMENT ****WATCH THE SYMBOL TABLE. REAL MEANS FLOATING(-POINT).
  144...     FUNCTION FIX IS CALLED IMPLICITLY WHEN A REAL EXPRESSION IS
  144...       ASSIGNED TO AN INTEGER CONSTANT. MIXED EXPRESSIONS FLOAT ANY
  144...       INTEGER PARTS AND YIELD A REAL RESULT****$
  144...    INTEGER C, H ,O,N,XNOM,CMIN,CMAX,HMIN,HMAX,HUNS,RWT,CLO,NO,OMIN,
  144...    NMIN,HONR$
  144...  INTEGER ARRAY HVAL(0..13)=(0,1,2,3,2,1,0,-1,-2,-3,-2,-1,0,1)$
  144... COMMENT SEE LEDERBERG AND WIGHTMAN, ANAL. CHEM., 36,2365(1964).
  144... THIS CORRECTION ALSO TAKES A/C RADICALS AND PROTONATED SPECIES$
  146...    WC=12.0$ WH=1.00782522$ WN=14.003074$ WO=15.994915$
  154...    COMMENT WH IS USED TO SCALE MASSVALUES TO THE BASE H=1.000
  154... WATCH ARITHMETIC FOR PRECISION
  154...    ON THIS BASE, C N AND O HAVE A MASS DEFECT....
  154...    128.795,  132.514,   123.769   DALTONS PER UNIT DEFECT$
  156...     WRITE($$HDR)$
  156... FORMAT HDR (B16,'**CALCULATION OF MOLECULAR FORMULAS**',W7)$
  173...    NEXTR..
  175...   READ($$MASSINPUTS)$ INPUT MASSINPUTS(MASSVALUE,MASSERROR,CARBMIN)$
  206...   IF MASSERROR LEQ 0$ MASSERROR=0.0001$
  213... COMMENT FOR PROGRAM CHECKOUT ONLY.  REALISTIC ERROR SHOULD BE READ IN.
  213... TRY ALSO MASSERROR = PRECISION. MASSVALUE$
  217...   IF CARBMIN EQL 0 OR CARBMIN GTR 0.999999$ BEGIN
  224...    CMIN=CARBMIN$ GO TO CARBATOM$ END      $
  225... COMMENT LOOPS FOR ADDNL. ATOM TYPES SHOULD GO HERE.
  225... THEY SHOULD MODIFY VALUES USED IN INNER LOOPS$
  235...    CMIN = (CARBMIN.(MASSVALUE-MASSERROR))/12$ CARBATOM..
  236...    IF CMIN LSS 0$ CMIN=0$
  241...   WRITE($$INR,HD2)$ OUTPUT INR(MASSVALUE,MASSERROR,CMIN)$
  255...    FORMAT HD2 (W2,'CALCULATION FOR MASS= ',X11.4,' +/- ',X7.4,
  277...    *     CMIN GIVEN AS *,I3,W2,
  311...  'MASS VALUE      C   H   O   N UNS    P/C C',W2)$
  311...   IF MASSVALUE LEQ 0$GO TO EOJ$
  314... COMMENT PRESUMED GARBLE$
  314...   XNOM=0.9+MASSVALUE. 0.99888$
  321... COMMENT INTEGER TRUNCATION AFTER SCALING MASS X 14.0/CH2.
  321... THIS WILL COVER ALL REAL COMPOUNDS OF C,H,O,N WITH,E.G., LSS THAN
  321... 20 CARBOXYL GROUPS. IF OTHER NUCLIDES ARE KNOWN TO BE ABUNDANT,
  321... MORE CAREFUL ANALYSIS IS INDICATED. XNOM IS INTERPRETED AS THE NOMINAL
  321... MASS IF EACH ATOM HAD INTEGRAL MASS$
  321...    M1 = MASSVALUE/WH$
  325...    MR = MASSERROR/WH$
  331...  MD=XNOM-M1$
  336...    MDL=MAX(0,MD-MR)$
  345...    MDH=MD+MR$
  350... COMMENT RELATIVE MASSDEFECTS, MEAN,LOW,HIGH$
  353...     XHI=MASSVALUE+MASSERROR$ XLO=MASSVALUE-MASSERROR$
  356...     XHI=MIN(XHI,3.WH+(XNOM-3).(1.001118 ))$
  373... COMMENT MAX AS IF ALKANE$
  405...     XLO=MAX(XLO,(WO.XNOM)/16)$
  406... COMMENT MIN IF ONLY OXYGEN. A GOOD POINT TO INPUT OTHER RESTRICTIONS$
  427...    HMAX=MIN((XNOM-12.CMIN), (HVAL(MOD(XNOM,14))+2.(XNOM/14)),
  452...    FIX(XNOM-123.76 . MDL))$
  454...    COMMENT DERIVED FROM (1).. RESIDUES AFTER CMIN ALLOCATED,
  454...   (2).. RESIDUES AFTER CONSIDERING MAX POSSIBLE CH2'S,
  454...   (3).. ALLOCATION NEEDED TO MAKE UP MASS FRACTION$
  460...    HMIN=XNOM-2.(FIX(132.52 .MDH )/2)$
  464...     IF HMIN LSS 0$ HMIN=MOD(XNOM,2)$
  473...     FOR H=(HMIN,2,HMAX)$ BEGIN
  504...     HF = H(0.00782522)$
  513...     NMIN=(XLO-XNOM-HF).(325.309)$
  523... COMMENT 1/(WN-14)$
  527...     OMIN = (XHI-XNOM-HF).(-196.656)$
  534... COMMENT 1/(WO-16)$
  542...     OMIN=MAX(0,OMIN)$ NMIN=MAX(0,NMIN)$
  550...     HONR=XNOM-(H+16.OMIN+14.NMIN)$ IF HONR LSS 0$ GO TO MOREH$
  565...     CLO =(15.H-XNOM-16.OMIN-42)/16$
  600...    COMMENT WORKING CMINIM BY NEED TO BIND H'S$
  605...     CLO=MAX(CLO,CMIN)$
  606...    CMAX= HONR/12$
  614... COMMENT PUT ADD'L DATA RESTRICTIONS HERE$
  625...     FOR C=(CLO,1,CMAX)$ BEGIN
  625... COMMENT ALGOL FOR LOOP... SKIPS IF CLO GTR CHAX$
  631...     RWT=XNUM-H-12.C$
  640...     NO=RWT/14$ O=(MOD(RWT,14)/2)$
  654...     IF O LSS OMIN$ BEGIN O=7.(OMIN/7)+MOD(O,7)$ IF O LSS OMIN$
  676...     O=O+7 $
  703...     N=(RWT-16.O)/14 $ GO NTEST $ END $
  715...     N=NO-O$
  720... NTEST..
  722...     IF N LSS NMIN$ GO TO NOGO$
  724...     GO TO TESTPQRS$
  725...     AGAIN..IF N GEQ 8$ BEGIN N=N-8$ O=O+7$ GO TO TESTPQRS END $
  740...     GO TO NOGO$
  741...     TESTPQRS..
  761...     XFL= C.WC +N.WN +O.WO +H.WH$
  770...     HUNS=N+2.C+2-H$  PCC=(1200.C)/XFL$
 1005...     IF XFL GEQ XLO AND XFL LEQ XHI AND HUNS GEQ -1$
 1017...     WRITE($$ASNWRS,XFF)$
 1024...     FORMAT XFF(X10.5,' = ', 5I3, X12.2,W)$
 1035...    OUTPUT ASNWRS(XFL,C,H,O,N,HUNS,PCC)$
 1056...     GO TO AGAIN$
 1057...   NOGO..END$
 1060...   MOREH..
 1060...   END$
 1061...     EOJ..
 1076...           PRINTOUT('...END OF JOB... READ NEW DATA')$
 1101...    GO TO NEXTR$
 1102...    FINISH$
RAM ENDS AT      1103
AMS END AT      11200
LES BEGIN AT   77336
```

```
!LOGIN SU010 JC093
 $OK LOG ON 28  10:10.4 03/08/66
LOAD MOLS
$LOAD 28
GO

$MSG IN.
                **CALCULATION OF MOLECULAR FORMULAS**
MASSVALU = ? 660.7512
MASSERRO = ? 0
 CARBMIN = ? 0
MASS NUMBER     C   H   O   N  UNS   %C
 660.75120 =   47 96   0   0   0    85.36
...EOJ...


MASSVALU = ? !TIME

$CLOCK TIME IS 10:11.
!QUIT

$MSG IN.
```

```
LOGIN SU010 JG093
 $OK LOG ON 23  10:02.8 03/08/66
LOAD MOLS
$LOAD 23
GO

$MSG IN.
                    **CALCULATION OF MOLECULAR FORMULAS**
MASSVALU = ? 660.7512
MASSERRO = ? 0
 CARBMIN = ? 0
MASS NUMBER     C  H   O   N  UNS   %C
 660.75120 =   47 96   0   0   0    85.36
...EOJ...

MASSVALU = ? !TIME

$CLOCK TIME IS 10:09.



!LOGIN SU010 JG093
$OK LOG ON 29  17:57.3  01/13/65
!LOAD MOLFORM
$LOAD 29
!GO

                    **CALCULATION OF MOLECULAR FORMULAS**
MASSVALU = ? 207.05411
MASSERRO = ?    .003
 CARBMIN = ?   0.5
MASS NUMBER     C  H   O   N UNS   %C
 207.05449 =   11  5   0   5 24    63.75
 207.05182 =    8  7   3   4 15    46.37
 207.05584 =   13  7   1   2 23    75.34
 207.05316 =   10  9   4   1 14    57.96
...EOJ...

!TIME
$CLOCK TIME IS 17:58
!QUIT
$MSG IN



                    **MOLECULAR FORMULAS::PARENT MOLECULAR IONS
MASSVALU = ? 963.6599
MASSERRO = ?    .0030
 CARBMIN = ? 20
  NMINIM = ?  7
  OMINIM = ?  7
MASS NUMBER     C  H   O   N UNS    %N
 963.65714 =   27 77   8  31 10    45.05
 963.65982 =   31 81  10  25  8    36.33
 963.66251 =   35 85  12  19  6    27.61
 963.65932 =   46 85   9  13 22    18.89
 963.65799 =   45 89  13   9 12    13.08
 963.66201 =   50 89  11   7 20    10.17
...EOJ...
```