A Data Structure for Rapid Mass Spectral Searching

Daniel L. Sweeney
MathSpec, Inc., 1314 N. Highland Avenue, Arlington Heights, Illinois 60004, United States

The combination of partitioning and systematic bond disconnection has been used to identify compounds from accurate-mass fragmentation data. This combination is very effective in excluding wrong answers that occur by chance. However, both processes are CPU intensive. This paper describes a novel data structure for representing molecules in a computer readable format that is conducive to very rapid mass spectral searching while still retaining the advantages of partitioning and systematic bond disconnection.

Please cite this article as: D. L. Sweeney, A Data Structure for Rapid Mass Spectral Searching, Mass Spectrom (Tokyo) 2014; 3(3): S0035; DOI: 10.5702/massspectrometry.S0035

Keywords: partitioning, identification, mass spectrometry, molecular structure, systematic bond-disconnection

INTRODUCTION

Identifying known compounds using mass spectrometry

The classical approach for identifying compounds from their mass spectral data is library matching. A mass spectral library is a computer medium containing a summary of the fragment masses and intensities of a large number of compounds that have been previously analyzed by mass spectrometry. In library matching, a search algorithm is used to compare the spectrum of an unknown compound to the fragment masses and intensities of all of the compounds in the library. A list of compounds in the library that best match the unknown compound is then produced. Library matching is especially useful for EI (electron ionization) spectra, because vast libraries of EI spectra exist—the combined NIST and Wiley EI libraries contain hundreds of thousands of spectra. Today only relatively small CID-type (collision-induced dissociation) mass spectral libraries exist, although public repositories of mass spectral data such as MassBank have been instituted and these databases are rapidly growing.

Data structures for representing structures of small molecules

The most widely used data structure for representing small molecules in a computer readable format is molfile format. This data structure consists of a computerized memory device in which molecules are represented with some header information (e.g. ID number), followed by the connection table (a listing of the one-to-one connections of the atoms that make up the molecule), and then sections of miscellaneous information.

Alternative computer compatible formats for representing molecular structures include InChI, SMILES, ASN1, and XML type data structures. These computer compatible representations will here be called molecular data structures.

Identifying compounds by comparing mass spectral fragmentation data to molecular structures

One advantage of searching a library of spectra is that library searches are very fast. Along this line, Hill et al. used commercial software (Mass Frontier) that predicts mass spectral fragments for a given chemical structure. They then constructed pseudo-fragmentation spectra of some compounds using these computed masses of the predicted fragments. They were then able to search mass spectral data of some known compounds against these computationally derived “spectra” of multiple compounds. This is essentially library searching. However, it appears that many more fragments are predicted than actually observed and improvements in the predictive software would be needed to make this approach more practicable.

Systematic bond-disconnection has been used to assign accurate-mass fragments to known compounds. Breakable bonds in a molecule were assigned a penalty score based on the likelihood that the bond will break. The rules to determine the penalty were much simpler and fewer than the rules used by the predictive software described previously. The bonds were then systematically broken, up to four at a time, and the masses and elemental compositions of the resulting pieces were found. Redundant masses and compositions were then removed. The masses of the fragment ions, obtained from the mass spectral data, were then compared to the calculated masses taking into account that the mass may differ by the number of hydrogens lost or gained in forming the fragment ion. If multiple pieces had the same mass and formula, the corresponding partial structures would be displayed.

This approach has been applied to the assignment of fragment ions observed in a mass spectrum and for me-
tabolite identification by comparison to the parent drug; the software is called "MassFragment." According to Waters, MassFragment assigns structures to observed fragment ions of small molecule compounds, drugs, and/or metabolites by systematic bond disconnection of the precursor structure instead of the traditional rule-based approach.

Sweeney described in great detail a process for deriving modular structures directly from CID-type mass spectral data; this process will herein be called partitioning.\(^4\) Mathematically, a partition is a set of integers that sum up to another integer.\(^5\) In mass spectrometry terms, partitioning is finding a set or sets of integral masses that sum up to the molecular weight of a compound and also account for many of the observed fragment ions of that compound.

The fragmentation of an organic compound in a mass spectrometer is not a random breaking of bonds; the breaking of a select group of bonds of the unknown organic compound yielding complementary subfragments can often account for most of the observed mass spectral fragments. This is the underlying principle of partitioning. Most organic compounds can therefore be represented in the form of unbreakable subfragments, of known elemental composition, joined together by breakable bonds. Modular structures basically show how mass spectral fragments may be related to each other within a whole molecule.

Each modular structure has a molecular formula. The fragment ions are viewed as different sets of contiguous subfragments; each subfragment has an elemental composition that is complementary to all of the other subfragments comprising the modular structure. For example, if the elemental composition of the whole molecule has only one sulfur atom, then assigning that sulfur atom to one particular subfragment will preclude all of other subfragments from having a sulfur atom.

To illustrate modular structures and subfragments, the Challenge 5 data was analyzed using an Excel Add-In that derives modular structures from accurate-mass fragmentation data using this partitioning approach. The results for the highest scoring partition are shown in an Excel worksheet (Fig. 1). Row 16 in the chart on the top right is a cylinder that represents the modular structure. The calculated masses (in daltons) of the subfragments are in row 23 (columns D, E, and F). Elemental compositions corresponding to isoprothiolane (Fig. 2) are in row 37; elemental compositions corresponding to diisopropyl dimethyl diphosphate (Fig. 2) are found in row 39.

The neutralized fragment masses (subtracting the mass of a proton from the experimental fragment ion data) are in column B rows 11 to 16. This partition and modular structure can account for the four non-zero masses in column C rows 11 to 16. With 3 subfragments, there are seven possible combinations: SubFragment1 \((42.0470)\), SubFragment2 \((187.9600)\), SubFragment3 \((60.0576)\), SubFragment1 + SubFragment2 \((230.0070)\), SubFragment1 + SubFragment3 \((102.1046)\), SubFragment2 + SubFragment3 \((248.0176)\), and SubFragment1 + SubFragment2 + SubFragment3 \((290.0646)\). However, one combination SubFragment1 + SubFragment3 cannot exist (without a rearrangement) because these two subfragments are not connected together in the modular structure. In addition, fragment ions corresponding to both SubFragment1 \((42.0470)\) and SubFragment3 \((60.0576)\) were not observed experimentally.

The experimental accurate-mass fragmentation data thus can be used to produce these modular structures that are composed of complementary subfragments. In an analogous manner, using systematic bond disconnection, a molecular structure can be partitioned into sets of complementary subgroups. While the hydrogen atom count in the subfragments may not equal the hydrogen atom count in the subgroups, the heavy atom compositions will be identical.
Based on a combination of systematic bond disconnection and partitioning, a commercial software program was launched in December 2006 to search the MDL® (now Accelerys) Available Chemicals Directory (Rational Numbers® FragSearch) with accurate-mass mass spectral data for the purpose of identifying unknown compounds.6) Rational Numbers® FragSearch software was comprised of a data processing means and four other major components. First, computerized molecular structures were represented in an abbreviated version of molfile format. Second, the mass spectral data of the unknown compound was analyzed by the data processing means and converted into plausible modular structures, connected groups of subfragments of known elemental composition. Third, all computerized molecular structures in the database having a molecular weight similar to the unknown compound were broken by systematic bond disconnection into complementary subgroups (connected groups of atoms that together with the other subgroups comprise a whole molecule; each heavy atom in a molecule can only be found in one subgroup). These connected subgroups are analogous to the subfragments in the modular structures derived from mass spectral fragmentation data by partitioning. Fourth, the heavy atom compositions of the subgroups and the subfragments were then compared using the data processing means.

Instead of using fragmentation rules to improve the selectivity of systematic bond disconnection, the partitioning process is used. Every search required that the molecules in the database with masses corresponding to the unknown compound must be broken by “systematic bond disconnection” for comparison with each of the possible modular structures of the unknown. Partitioning and systematic bond disconnection, required for searching this way, are both very CPU intensive, especially for larger molecules with more bonds and more partitions. The original version of Rational Numbers Search ran on a Mac mini and the process was very slow, often taking hours. To provide faster results, a much more powerful data processing means than a single workstation was later employed. The Rational Numbers® FragSearch application was provided as an application on the Sun Grid Compute Utility (SGCU, later called the Sun Cloud).7,8) This utility allowed searches to be conducted in parallel on multiple 64-bit Opteron processors. The Sun Grid Compute Utility, never fully implemented by Sun, was abandoned by Sun Microsystems in October 2008. The utility of this approach appeared to be constrained by a lack of available and easy-to-use high throughput CPU resources.

From a different perspective, there are many ways to represent chemical structures on a computer, but no present data structure is really conducive to rapid mass spectral searching. This paper describes a data structure for representing molecules that is conducive to very rapid mass
A Data Structure For Rapid Mass Spectral Searching

The representation of isoprothiolane in this data structure and its accurate-mass data will be used here for illustrative purposes. The Challenge 5 dataset (isoprothiolane) is very accurate and therefore can also be used to demonstrate the effect of varying the mass error window (MaxDefect parameter).

THE DATA STRUCTURE

An example

Essentially, isoprothiolane can be represented by a set of exact masses of partitions of contiguous subgroups, and a unique ID number. Table 1 lists all of the 3-subgroup set of partitions that represent isoprothiolane (PubChem ID# 39681) and diisopropyl dimethyl diphosphate (PubChem ID# 218517). In this 3-subgroup example, each row represents a different and unique partition of a molecule. The first three columns are the exact masses (in units of tenths of millidaltons) of each complementary subgroup (SubGrp) designated as subgroups A to C. These three subgroup masses sum to the whole mass: the combination of the three subgroups in the seventh column. Thus each row is by definition a mathematical partition of the exact mass of a molecule. The middle columns are all of the combinations of subgroups, and the last column is the PubChem Compound ID number. If two subgroups are not connected to each other in the molecule, that combination is given a mass of 0. The masses of the bolded and italicized subgroups and combinations of subgroups in Table 1 correspond to the subfragment masses and combinations of subfragment masses in column C rows 11 to 16 (Fig. 1). Note that the masses of subfragments and subgroups may differ by the mass of an integral number of hydrogen atoms in addition to the ex-

Table 1. The three subgroup (SubGrp) partitions of isoprothiolane (PubChem ID# 39681) and diisopropyl dimethyl diphosphate (PubChem ID# 218517).

| SubGrp A | SubGrp B | SubGrp C | A+B | A+C | B+C | A+B+C | PubChem ID |
|----------|----------|----------|-----|-----|-----|-------|------------|
| 159950   | 430548   | 230150   | 590498| 2470100| 0     | 2900647  | 39681      |
| 590497   | 279950   | 2030200  | 870447| 0    | 2310150| 0      | 2900647   | 39681      |
| 218065   | 600034   | 430548   | 2470100| 2620335| 0     | 2900647  | 39681      |
| 202938   | 280313   | 590497   | 2310151| 2620335| 0     | 2900647  | 39681      |
| 759442   | 280313   | 1860892  | 1039755| 2620334| 0     | 2900647  | 39681      |
| 1749888  | 280313   | 870446   | 2030201| 2620334| 0     | 2900647  | 39681      |
| 319722   | 2150378  | 430548   | 2470100| 2620335| 0     | 2900647  | 39681      |
| 439899   | 430548   | 2030200  | 870447   | 2620334| 0     | 2900647  | 39681      |

Masses are in tenths of millidaltons. The partitions that are discussed in the text are italicized and bolded. The number 0 indicates that the two subgroups indicated are not connected in the molecule.
Experimental error in the subfragments.
Some compounds have subgroups of identical mass. For example, PubChem ID # 60807 (Fig. 2) has four identical butyrate groups and systematic bond disconnection will generate a considerable number of identical partitions for compounds like this. However, only one of these identical sets is stored in the data structure. This cuts down on the number of partitions and eliminates essentially redundant answers that otherwise would arise. In this data structure, each of the partitions in the set of partitions representing a given compound are unique.

Sets with various numbers of subgroup masses are generated. For example, each molecule could be broken into sets of 2 subgroups, 3 subgroups, 4 subgroups, 5 subgroups, etc. The present data structure has each of the 250,000 compounds partitioned into sets of 2 subgroups, 3 subgroups, and 4 subgroups. For simplicity, only the 3-subgroup partitions are described in this report.

A unique feature of this representation of molecules is that the entire data structure is made up of numbers. Thus the numerical data structure can be searched rapidly against the numerical mass spectral data with typical search times of less than two seconds. There is no time consuming interconversions of numbers and atoms.

How the data structures are made
Essentially, systematic bond disconnection is used to generate the data structure, but only disconnections of multiple bonds that yield partitions of the molecular weight are collected in the data structure.

One consideration in using molecular structures rather than libraries is that many common compounds may be salts or quaternary amines. So, as an example, prior to doing systematic bond disconnection, the antihistamine chlorpheniramine maleate would be converted into chlorpheniramine, the maleate counter ion being removed from the structure. In addition, mono-quaternary amines are stored at an exact mass one proton lower than theoretical and without the corresponding counterion. These operations are done prior to the systematic bond disconnection.

Also prior to doing systematic bond disconnection, some bonds of compounds that are being processed are “locked.” There is no attempt to score each bond on how likely that particular bond may break. Either a bond can break or not break (locked). For example, it is very unusual for a benzene ring to fragment under CID fragmentation conditions—unless one of the carbon atoms of the ring is attached to an activating group such as an oxygen. Therefore the ring bonds in most molecules containing benzene and naphthalene rings are locked. In addition, the bonds of aliphatic hydrocarbon chains are also locked. Triple bonds are locked. The consequences of locking some bonds are that the number of partitions is fewer and searching is therefore faster.

Based on the representation of a molecular structure in molfile format, systematic bond disconnection is then applied to breakable bonds in the structure and the structure is broken into pieces. A 3-subgroup representation will be used to illustrate how the representations are made. The objective therefore is to break the molecule into three pieces in which each heavy atom (and its attached hydrogens) is found in only one of the three pieces; the pieces are complementary. To break a molecule into three pieces, at least two bonds must be broken simultaneously. If cyclic moieties are present, then it might be necessary to break three, four, or more bonds to get three pieces. To generate representations of 3 subgroups, the systematic bond disconnection is therefore applied to combinations of all breakable bonds, taking 2, 3, or 4 bonds at a time. Often the wrong number of pieces (2, 4, 5 etc.) might be generated; these are rejected. When 3 pieces that are partitions of complementary subgroups that comprise the whole molecule are generated, the exact mass of each complementary subgroup is then calculated. Because the exact mass of each heavy atom in a subgroup can only be found in one subgroup of exact masses, each molecule is “partitioned” into exact masses of subgroups. The number of disconnected bonds needed to generate each individual subgroup and the PubChem ID number are also stored with each row. After generating the partitions, some compounds having subgroups with the same elemental composition will have some identical partitions. However, only one of these identical sets is stored in the data structure and the redundant partitions are removed.

Basically, the very CPU-intensive processes of systematic bond disconnection and removal of redundant partitions that were required to make this data structure, have now been completely separated from the actual search process.

The number of sets of subgroups required to represent a chemical compound increases with the molecular weight and number of bonds in the compound. However, since there are fewer higher mass compounds, there is not much difference in the total number of sets of masses as the molecular weight increases. Thus search times do not vary significantly with the molecular weight of the unknowns.

SEARCHING

Preprocessing the experimental mass spectral data prior to searching
Two different input formats can be selected, based on the instrumentation available and how the mass spectral data is collected. In-source CID or “All Ions” or MS² data usually consist of a low energy dataset used to determine the molecular weight of the unknown compound and a high energy dataset used to obtain the fragment ions. For MS/MS data, only one dataset is needed but it must contain at least some of the precursor ion (>1.5% relative intensity). Multiple datasets (e.g. MS/MS at three different collision energies) can be combined—but at least one dataset must have the precursor ion present.

The data is first neutralized by adding the mass of a proton to negative mode data or by subtracting the mass of a proton from positive ion data. The dataset is also desoxygenated to obtain monoisotopic data. Less intense (<1.5% relative intensity) fragment ions are dropped. A “coverage” value for each fragment is then calculated which is based roughly on the square root of the experimentally measured relative intensities, so that the more intense ions are given less weight.

This preprocessing is done as part of the actual search and the intermediate files are not saved. However, the original data and preprocessed dataset for isoprothiolane (Challenge 5) are shown in Table 2 for illustration.

A parameter known as MaxDefect is of critical importance. This is the maximum error in millidaltons by which the
fragment mass data is expected to vary from the theoretical masses of subgroups and combinations of subgroups of compounds in the data structure. For this approach to be practicable, it is necessary that the experimental accuracy of the mass spectrometer laboratory be assessed using known compounds under typical conditions. The MaxDefect parameter should then be set at the lowest value possible, but where about 95% of the fragment ions will be inside this mass error window. This key parameter should also not exceed 5 millidaltons, a value most modern accurate-mass instruments can achieve.

### Table 2. The positive MS/MS and MS data for Challenge 5.

| Mass (Da) | Intensity | Mass (Da) | Intensity | Mass (Da) | Relative intensity |
|----------|-----------|-----------|-----------|-----------|-------------------|
| 53.19455 | 7279.5    | 282.13838 | 46832.2   | 230.006912| 100               |
| 56.519   | 11297.9   | 284.29474 | 106511.3  | 187.96059 | 15                |
| 59.0292  | 6424.9    | 285.13582 | 57472.7   | 248.01768 | 7                 |
| 61.52185 | 3119.3    | 286.12539 | 63430     | 171.96530 | 2                 |
| 66.21554 | 9003.4    | 286.60496 | 69952.8   | 205.97077 | 2                 |
| 105.98727| 5448.5    | 289.0217  | 49916     | 290.06436 | 0                 |
| 144.97753| 8148.9    | 291.05534 | 88112     |           |                   |
| 149.02326| 54437.2   | 291.07187 | 1631953   |           |                   |
| 167.03379| 9225.8    | 291.08559 | 89512.7   |           |                   |
| 172.96156| 5905.4    | 291.12183 | 26276.6   |           |                   |
| 172.96425| 6294.7    | 292.07538 | 1807586.5 |           |                   |
| 172.97256| 680091.4  | 292.12516 | 105232.8  |           |                   |
| 181.7542 | 12373     |           |           |           |                   |
| 188.96743| 3946115   | 293.06762 | 1275843.6 |           |                   |
| 188.97624| 11738.5   | 293.13305 | 177858.9  |           |                   |
| 189.97065| 5318.5    | 293.17853 | 65854.3   |           |                   |
| 190.98309| 3639.9    | 293.6348  | 38267     |           |                   |
| 196.61872| 8605.0    | 294.07105 | 138577.7  |           |                   |
| 206.97805| 641602.8  | 295.1943  | 106376.2  |           |                   |
| 221.29375| 28732.2   | 296.50946 | 74491.2   |           |                   |
| 230.9403 | 3193.6    | 298.00815 | 94168.3   |           |                   |
| 230.94924| 3996.4    | 299.00573 | 31897.1   |           |                   |
| 230.95463| 4889.5    | 300.62076 | 46030.6   |           |                   |
| 230.96262| 10525.9   | 301.14088 | 27341.3   |           |                   |
| 230.97549| 13974.3   | 302.05649 | 28939.6   |           |                   |
| 230.98625| 19945.3   | 306.14084 | 31877.7   |           |                   |
| 231.0147 | 24729928  | 308.15417 | 97248.9   |           |                   |
| 231.03082| 42643.1   | 310.0453  | 986685    |           |                   |
| 231.03898| 26317.9   | 310.16959 | 79298.4   |           |                   |
| 231.04435| 16417.1   | 310.54695 | 251875.1  |           |                   |
| 231.04679| 15205     | 311.0428  | 158802    |           |                   |
| 231.04955| 14041.7   | 311.15159 | 89567.9   |           |                   |
| 231.05222| 12909     | 313.05297 | 5782337.5 |           |                   |
| 231.05468| 12741.2   | 313.16674 | 94968.7   |           |                   |
| 231.05741| 10412.8   | 314.05744 | 627548.1  |           |                   |
| 231.064   | 10302     | 315.04963 | 429885.3  |           |                   |
| 231.07058| 5211.2    | 316.05289 | 41320.9   |           |                   |
| 231.07626| 3536.7    | 317.53303 | 513540.7  |           |                   |
| 231.0789  | 3695.3    | 318.03461 | 131164.3  |           |                   |
| 231.08361| 2897.6    | 318.53036 | 69231.1   |           |                   |
| 231.16697| 4324.6    | 319.03164 | 482914.2  |           |                   |
| 231.42237| 6414      | 319.05066 | 63130.3   |           |                   |
| 232.01742| 58584.2   | 319.53336 | 119692.1  |           |                   |
| 241.38949| 11565.5   | 320.02933 | 235706.3  |           |                   |
| 249.02495| 1830133   | 320.53087 | 54968     |           |                   |
| 273.62957| 8133.7    | 321.02723 | 46913.9   |           |                   |
| 275.29267| 11667.1   | 329.02778 | 1691603.6 |           |                   |
| 279.15898| 40738.9   | 330.03143 | 185268.8  |           |                   |
| 291.0719 | 990182.8  | 331.02451 | 173465    |           |                   |
| 291.16289| 8535.7    | 338.34173 | 60044.1   |           |                   |

The pre-processing step removes adducts and minor fragment ions, determines the molecular weight for “All Ions” or MS² or CID-MS data, and also de-isotopes the data. The preprocessed masses have also been neutralized here by removing the mass of a proton. Masses are in daltons.

Searching the data structure for a matching compound in the data structure

The sets of partitions of exact masses of subgroups can be compared to the accurate masses of fragment ions generated on the mass spectrometer, while taking into account that the masses of the subgroups and combinations of subgroups will often differ from the mass of the corresponding fragment ion by some multiple of the exact mass of a hydrogen atom in addition to the experimental error. The basic process for searching is briefly described here.

First the partitions having the same mass as the unknown compound (“same mass” meaning within the MaxDefect er-
ror window) are stored in an array.

For three subgroups, the comparison is basically between one partition of masses of seven subgroups and combinations of subgroups being compared to the set of masses of neutralized fragment ions. Each neutralized fragment ion is compared sequentially to masses making up the partition. A match is found if the two masses are within the MaxDefect window.

Since the masses will often differ by the mass of an integral number of hydrogen atoms, the search process is somewhat more complicated. The seven masses of the 3-subgroup partition are sequentially shifted by the mass of a hydrogen atom. This hydrogen adjustment process is repeated until a maximum number of exact masses of hydrogen atoms is added or subtracted. The maximum number of hydrogen mass shifts allowed for each mass in a partition is a function of how many bonds were broken to generate that mass. After each hydrogen shift, each neutralized fragment ion is compared sequentially to masses making up the partition. The necessity of doing the hydrogen shifts explains why this searching approach cannot be done as a typical direct database search.

If the mass difference is within the MaxDefect window, the score for that partition is increased by the coverage value of that fragment ion. After the completion of the hydrogen shifts, the initial score (sum of the matching coverages) obtained is then adjusted by several parameters (mass accuracy, number of fragments matched, and the number of bonds that were broken to form the partition). The next partition of subgroups of exact mass is then compared to the neutralized fragment ions.

Partitioning of the mass spectral data to find the exact masses of subfragments and to remove trivial linked partitions is no longer needed. Previously, trivial linked partitions were removed during the partitioning process. If two subfragments are linked, this basically means that two subfragments are being used to do the work of one subfragment. By using flags during the searching process, linked partitions can be detected and removed; searching can therefore be done without prior partitioning of the mass spectral data, eliminating the second CPU intensive process from the search process.

Searching the data structure for similar compounds

This data structure is amenable to “Like” searching. For example, suppose the saturated 5-membered ring of isoprothiolane was replaced with an identical ring except that it had one double bond (e.g. PubChem ID# 4006, Fig. 2). This would therefore reduce the molecular weight of the whole compound by the mass of two hydrogens. We would also expect that this compound, if partitioned and present in the database, would have partitions that were almost identical with the partitions of isoprothiolane—except that one subgroup would be decreased by the mass of two hydrogens. In addition, every combination containing that subgroup (and only combinations with that subgroup) would be expected to shift by that same mass relative to isoprothiolane. Essentially, the accurate-mass data for isoprothiolane could be used to find compounds in the database that are similar to isoprothiolane.

In the present embodiment of this approach, multiple common mass shifts are checked. This includes loss or gain of masses corresponding to H2, CH2, C2H4, C3H6, C4H8, O, and H2O.

To put this approach into practice, after the molecular weight of the unknown compound is determined in preprocessing, partitions with an exact mass within the MaxDefect of the compound’s apparent mass as well as all compounds with an exact mass adjusted by these common mass shifts are stored in the initial array of partitions to be tested. In addition to the hydrogen adjustment previously described, now the masses of each subgroup (one at a time) are adjusted by the mass shift being tested.

Because more compounds in the database are being checked and because there are multiple subgroup positions possible for the mass shift, the search time increases compared to the previous search strategy described, but it is still generally less than 30 seconds even with 4-subgroup searching.

Incorporating MS" data

MS" data, if available, can be used to improve the selectivity. For example, assume a fragment ion is composed of contiguous subgroups A and B. Then, when this fragment ion is fragmented further, it cannot produce any product ion containing subgroup C (assuming that the masses of the subgroups are not equal). The availability and use of MS" data in this way can make the search process more selective.

RESULTS

Searching for an exact compound match

A typical result file is shown in Fig. 3. The top section is a listing of compounds that appear to match the fragment data; the lower section is a listing of compounds that could be considered a molecular weight match based on the MaxDefect (here 5 millidaltons). The top section has scores in the left column. The scores are all relative and based only on the mass spectral data. The column on the right side is the PubChem ID which is also a hyperlink to the PubChem Compound database.

The number of synonyms is listed for each compound. This is an indication of the relevance of the compound, analogous to the number of references used by Little et al. A compound with a large number of synonyms is usually in widespread use and more likely to occur in a sample than a compound with only a few synonyms. The number of synonyms does not affect the scoring.

The lower section lists the predicted ratios of the first four isotopologues for each compound. The isotope ratios are not used in scoring at this time, because it was found that poor isotope ratio data often excluded the correct answer. This listing of isotope ratios allows for a convenient visual comparison.

Searches are normally done on 2, 3, and 4 subgroup partitions and then the results are pooled. For the sake of simplicity here, the Challenge 5 data was analyzed using only 3-subgroup data structures with a MaxDefect of 5.0, 2.5, or 1.0 millidaltons. The results are shown at the top of Table 3. As the error window gets smaller, more compounds are excluded but in this case there was no improvement in going from 2.5 to 1.0 millidaltons.

Comparing the structures of isoprothiolane (PubChem
ID# 39681) and diisopropyl dimethyl diphosphate (PubChem ID# 218517), both compounds are diisopropyl esters. However, the remaining parts of the two compounds are completely different. Subgroups of two partitions are compared below, taken from Table 1, with masses in units of tenths of millidaltons. Because of the accuracy of the data, the large isoprothiolane subgroup is much closer in mass to the experimental value of the neutralized fragment at 187.960159 daltons (Table 2) and therefore isoprothiolane receives a higher score. However, this does illustrate that it is very common that the incorrect matches obtained are completely unrelated to the compound that generated the fragmentation data.

\[
1879603 \quad 430548 \quad 590497 \quad 39681
\]

Isotope ratios, if trustworthy, can be very helpful and here favored isoprothiolane. The ratio

100 11 8 1 was calculated for the Challenge 5 experimental data,
100 15 11 1 was calculated for isoprothiolane, and
100 9 2 0 was calculated for diisopropyl dimethyl diphosphate.

**Searching for similar compounds**

The Challenge 5 data was analyzed using the 3-subgroup data structure with MaxDefects of 5.0, 2.5, or 1.0 millidaltons. These results are shown at the bottom of Table 3. As the error window decreases, more compounds are excluded as expected.

Two very similar compounds were evident: PubChem ID# 4006 and PubChem ID# 93275 (Fig. 2). The scores of these two compounds and isoprothiolane were virtually identical. For each compound two relevant partitions are compared in Table 4.

**DISCUSSION**

The major advantage is speed. The slow process of systematic bond disconnection is no longer part of the actual search process. In addition, there is no need to convert back and forth between elemental compositions and masses. Both the representation of chemical structures and the fragmentation data are formatted as numbers.

In addition, there is no need to do prior partitioning of the mass spectral data. Partitioning, through systematic bond disconnection, is only done on the molecular struc-
tures. Previously, partitioning was done on the mass spectral data and one perceived advantage was that partitioning was able to eliminate "linked partitions." However, as mentioned previously, by using flags it is possible to eliminate linked partitions without actually doing the CPU-intensive partitioning on the mass spectral data.

A second advantage is simplicity. There are very few rules with respect to bond breaking. It is difficult to predict how a given compound will fragment in a mass spectrometer even with 20,000 rules. Here, there is no need to score how likely a given bond is to break; bonds are classified only as locked or breakable. The simplicity and uniformity of this data structure is also well suited for GPU processing with CUDA and similar multi-CPU approaches to high-throughput parallel computing.

**CONCLUSION**

The overall strategy is to limit the compounds in the data structure to about 250,000 common compounds, which is slightly larger than the number of unique compounds in typical GC/MS libraries. At this size and with accurate-mass accuracy of about 5 millidaltons, this approach is capable of identifying common compounds from accurate-mass fragmentation data while keeping CPU search time fast enough so that the compounds could potentially be identified as rapidly as data is generated. Currently, the major hurdle is developing a convenient means for transmitting mass spectral data into a format that can be readily searched against this data structure.

The proliferation of designer drugs is a growing medical and social problem in developed countries. Generally, minor modifications are made to the molecular structures of some drugs (such as sildenafil) to evade targeted mass spectral approaches. The similarity searching approach described here may be very useful for detecting and identifying designer drugs.

**Acknowledgements**

The author appreciates the efforts of Dr. Takaaki Nishioka and his associates for organizing and conducting the 2013 CASMI challenge.

**REFERENCES**

1) D. W. Hill, T. M. Kertesz, D. Fontaine, R. Friedman, D. F. Grant. Mass spectral metabonomics beyond elemental formula: Chemical database querying by matching experimental with computational fragmentation spectra. *Anal. Chem.* 80: 5574–5582, 2008.

2) I. A. Watson, A. Mahoui, D. C. Duckworth, D. A. Peake. A strategy for structure confirmation of drug molecules via automated matching of structures with exact mass MS/MS spectra. Proceedings of the 53rd ASMS Conference on Mass Spectrometry, June 5–9, 2005, San Antonio, TX.

3) A. Hill, R. Mortishire-Smith. Automated assignment of high-resolution collisionally activated dissociation mass spectra using a systematic bond disconnection approach. *Rapid Commun. Mass Spectrom.* 19: 3111–3118, 2005.

4) D. L. Sweeney. Small molecules as mathematical partitions. *Anal. Chem.* 75: 5362–5373, 2003.

5) http://en.wikipedia.org/wiki/Partition_of_a_set

6) D. L. Sweeney. A systematic computational approach for identifying small molecules from accurate-mass fragmentation data. *American Laboratory* News 39: 12–14, 2007.

7) http://en.wikipedia.org/wiki/Sun_Cloud.

8) http://www.ogf.org/OGF24/materials/1446/Ent10+-+Sub+Initiative.pdf

9) D. L. Sweeney. A representation of molecules as sets of masses of complementary subgroups and contiguous complementary subgroups. United States Patent Application 2011017161, published July 14, 2011.

10) R. Martin. "LC-MS²... All the Data, All the Time," http://www.waters.com/webassets/cms/library/docs/720001814en.pdf

11) M. Josefsson, M. Roman. "Rapid and Accurate Forensics Analysis using High Resolution Ions MS/MS," Agilent Application Note, http://www.chem.agilent.com/Library/applications/5991-2319EN.pdf

12) J. L. Little, A. J. Williams, A. Pshenichnov, V. Tkachenko. Identification of “known unknowns” utilizing accurate mass data and ChemSpider. *J. Am. Soc. Mass Spectrom.* 23: 179–185, 2012.