Accelerated Isotope Fine Structure Calculation Using Pruned Transition Trees

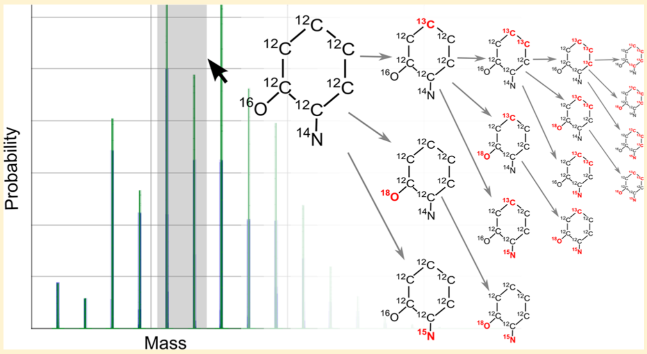
Martin Loos,*,[†,‡] Christian Gerber,[†] Francesco Corona,[§,∥] Juliane Hollender,[†,‡] and Heinz Singer[†]

[†]Eawag, Swiss Federal Institute for Aquatic Science and Technology, 8600 Dübendorf, Switzerland

[‡]Institute of Biogeochemistry and Pollutant Dynamics, ETH Zürich, 8092 Zürich, Switzerland

[§]Department of Information and Computer Science, Aalto University, 02150 Espoo, Finland

[∥]Department of Teleinformatics Engineering, Federal University of Ceará, Fortaleza, Ceará 60020-181, Brazil



ABSTRACT: A fast and memory-efficient calculation of theoretical isotope patterns is crucial for the routine interpretation of mass spectrometric data. For high-resolution experiments, calculations must procure the exact masses and probabilities of relevant isotopologues over a wide range of polyisotopic compounds, while pruning low-probable ones. Here, a novel albeit simple treelike structure is introduced to swiftly derive sets of relevant subisotopologues for each element in a molecule, which are then combined to the isotopologues of the full molecule. In contrast to existing approaches, transitions via single replacements of the most abundant isotope per element are used in separable tree branches to derive subisotopologues from each other. Moreover, the underlying transition trees prevent redundant replacements and permit the detection of the most probable isotopologue in a first phase. A relative threshold can then be exploited in a second parallelized phase for a precise prepruning of large fractions of the remaining subisotopologues. The gain in performance from such early pruning and the lower variation in the distortion of simulated data with use of relative rather than absolute thresholds were validated in a large-scale benchmark simulation, unprecedentedly comprising several thousand molecular formulas. Both the algorithm and a wealth of related features are freely available as R-package *enviPat* and as a user-friendly Web interface.

Effective calculation of isotopic patterns from molecular formulas is essential to understand mass spectrometric (MS) measurements. The simulated tuples of isotopologue masses and probabilities are required for restrictions during unknown identification,[1,2] targeted screening,[3] nontargeted signal grouping,[4] and the annotation of product ions in MS^n experiments,[5] among others. Herein, large batch calculations of candidate isotope patterns often need to be derived and compared, demanding computational speed and memory. On top, the measured molecules cover a wide array of polyisotopic elements and range from small molecules to polypeptides and proteins, typically with vast numbers of isotopologues per molecule even at lower masses.

Tackling the combinatorial complexity of simulating the measurable sets of such isotopologues has led to a plethora of approaches, not all of which scale well with increasing number of isotopes, elements, atoms or with instrument resolution.[6] These approaches can broadly be classified into two groups. One group either aggregates the individual isotopologues to nominal and center masses or directly samples the convolution of resolution-dependent peak shapes.[7] Although results may range within measurable accuracies, the underlying isotope fine structures are not fully resolved. This structure is needed for data interpretation or when convoluting a calculated set of isotopologues for different instrument resolutions. The latter can differ by orders of magnitude at a given mass and subsequently result in widely differing measurement signals.[8] Even more concerning, computational limitations are likely to arise for methods of this first group with the advent of ever higher resolutions.[9] Given that widespread time-of-flight and Orbitrap MS instrumentation already achieves resolutions well above 1 × 10^5, a comparison of highly resolved measurements with rapidly simulated isotopic features is central to, for example, complement monoisotopic mass information in molecular formula assignments. A second, albeit smaller, group of approaches has therefore been established, yielding isotope fine structures. These approaches are based on dynamic programming,[10] transitions in the multinomial distribution,[11−13] or most recently, multidimensional Fourier transforms.[9]

With isotopologue numbers easily exceeding computational resources, all these isotope fine structure approaches must prune their low-probability isotopologues. In some cases, such pruning is paid by imprecision for the resulting isotopologues[10]

Received: March 10, 2015
Accepted: April 30, 2015
Published: April 30, 2015

DOI: 10.1021/acs.analchem.5b00941
Anal. Chem. 2015, 87, 5738−5744

or risks a loss of sufficiently probable ones at subsequent calculation stages.[11] But even in the precise cases, pruning has been restricted to the late stage of combining subisotopologue information on the individual elements stripped from the molecule.[9,11,13] Large subsets of these subisotopologues have then been calculated in vain, whereas a pruning strategy at earlier stages has not been presented. The latter bears potential to significantly accelerate calculations. In addition, the named approaches base their pruning on absolute probability thresholds[9] or proportions of least probable isotopologues[10] or truncate negligible mass states on the basis of cumulative probabilities.[11,12] A relative probability threshold to be set as a fraction of the most dominant molecular isotopologue has been neither proposed nor evaluated—simply because this most probable isotopologue is not known in the first place. One can anticipate a relative pruning strategy to scale better with the widely differing dispersion of isotopologue probabilities among the analytes of interest.

On this background, we introduce a novel approach to compute isotope fine structures, augmenting the transition methodology introduced by Yergey[11] for directly calculating isotopologues from each other by single isotope replacements. Unlike his stepwise approach or the arrangements of isotopologues into levels of increasing mass proposed by Li et al.,[12,13] our method organizes transitions between isotopologues by gradually replacing the most abundant isotope of each element with variable subsets of less abundant ones. The underlying treelike structure takes advantage of the multinomial nature of isotopologue transitions by separating branches of decreasing probability from increasing ones. We explicitly address how this structure can be used to (a) generate all feasible isotopologues, (b) avoid redundant transitions, (c) swiftly find the most probable isotopologues, and (d) enable a very efficient pruning based on relative thresholds. We also elucidate the scaling properties of our approach with an unprecedented benchmark set of several thousand molecular formulas.

■ METHODS

Rationale. In general, the MS signal of a molecule can be calculated by (1) dividing the molecule into submolecules for each element, (2) deriving the isotopologues within these submolecules, (3) combining these subisotopologues to the exact isotopologue probabilities and masses of the full molecule, and (4) convoluting to measurable spectra at a given resolution by use of peak shape functions.[9,11,13] These steps will be exemplified for the sodium adduct $C_{20}H_8O_{10}Br_4S_2Na_1^+$ of the dye bromsulphthalein.

First, the molecule is divided into submolecules that contain only the n_k atoms of the kth of a total of $k_{\max}$ elements, that is, C, H, O, Br, S, and Na at $n_1 = 20$, $n_2 = 8$, $n_3 = 10$, $n_4 = 4$, $n_5 = 2$, and $n_6 = 1$, respectively. In turn, $n_{k,i}$ stands for the number of atoms composed of a certain isotope i in the kth submolecule; $a_{k,i}$ and $m_{k,i}$ are the natural abundance and mass of this isotope, respectively. Sorted by decreasing abundance, $i = 1$ denotes the monoisotopic variant and often coincides with the lightest isotope of an element in organic compounds. In the given example, $k = 3$ and $i = 1$ refers to ^{16}O. Similarly, $n_{3,1} = 10$ is the monoisotopic subisotopologue of oxygen.

For each submolecule, transitions between any two subisotopologues x and y that differ in only one isotope allow for a simple updating of the probability $P_{k,y}$ and mass $M_{k,y}$ of y from the probability $P_{k,x}$ and mass $M_{k,x}$ of x:[11,12]

$$P_{k,y} = P_{k,x}\frac{n_{k,i}}{n_{k,j}}\frac{a_{k,j}}{a_{k,i}} \tag{1}$$

$$M_{k,y} = M_{k,x} - m_{k,i} + m_{k,j} \tag{2}$$

Herein, i denotes the isotope in x that is replaced by isotope j to produce subisotopologue y, with $i \neq j$. The exact position of the replaced isotope in a submolecule is irrelevant; each subisotopologue may consist of several isotopic isomers. For example, the transition of $k = 5$, $i = 1$, and $j = 2$ is the replacement of one ^{32}S isotope by a ^{34}S isotope in the sulfur submolecule. Furthermore, each set of subisotopologues is initialized with one entry, having probability $P_{k,1} = a_{k,1}{}^{n_k}$ and mass $M_{k,1} = n_k m_{k,1}$, based on the most abundant isotope $i = 1$ only. Starting from this specific subisotopologue, transitions are recursively applied to devise the probabilities $P_k = \{P_{k,1}, ..., P_{k,b_k}\}$ and masses $M_{k,1} = \{M_{k,1}, ..., M_{k,b_k}\}$of a relevant set of 1, ..., b_k subisotopologues per submolecule k. Notably, different transitions can produce the same subisotopologue. For instance, the subisotopologue $^{33}S_1{}^{34}S_1$ can be formed by transition from both $^{32}S_1{}^{33}S_1$ and $^{32}S_1{}^{34}S_1$. A methodology to avoid such computational redundancies will be proposed in a later section on transition trees.

In the third step, subisotopologues from the different submolecules are combined via the $k_{\max}$-fold Cartesian products of their probability and mass sets:

$$P_1 \times \cdots \times P_{k_{\max}} = \{(P_{1,j_1}, ..., P_{k_{\max},j_{k_{\max}}}): P_{i,j} \in P_i\} \tag{3}$$

$$M_1 \times \cdots \times M_{k_{\max}} = \{(M_{1,j_1}, ..., M_{k_{\max},j_{k_{\max}}}): M_{i,j} \in M_i\} \tag{4}$$

The resulting $k_{\max}$-tuples of the product set contain all ordered combinations of subisotopologues from the different submolecules. To finally derive joint probabilities of the isotopologues of the full molecule, the entries within each tuple from eq 3 are multiplied. In contrast, entries within each tuple from eq 4 are summed to molecular isotopologue masses and corrected for electron masses if charged.

Notably, each molecule has a maximum of

$$b_{\text{tot}} = \prod_{k=1}^{k_{\max}} \binom{q_k + n_k - 1}{n_k} \tag{5}$$

isotopologues, where q_k is the total number of isotopes of the kth element.[10] The binomial coefficient defines the number of unordered combinations in which n_k atoms can be composed of q_k isotopes. Even for small molecules, b_{tot} can become prohibitively large and often contains a dominant fraction of isotopologues with negligible probability. Indeed, the small exemplary molecule has an isotope fine structure containing $b_{\text{tot}} \approx 6.2 \times 10^5$ isotopologues. However, over 99.9% of these isotopologues have probabilities at fractions of $<1 \times 10^{-5}$ of the most abundant isotopologue. Strategies to omit such low-abundance (sub)isotopologues at the level of both transitions and Cartesian product sets are therefore the subject of a later section on pruning.

Finally, peak shape functions of the remaining isotopologues are superimposed in a fourth step and scaled to relate to the isotopic envelopes of measured mass spectra at a certain instrument resolution R (Figure 1). Although this theoretical envelope is continuous, it is typically sampled at regular mass intervals to match the discretized measurement spectra. Further

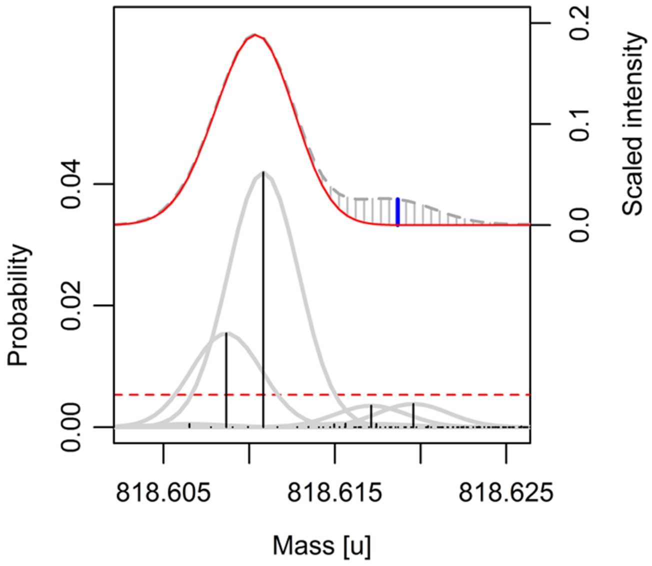


Figure 1. Superposition of Gaussian peak shapes (gray lines) of individual isotopologues (black bars) produces an isotopic envelope (gray dashed line) to be sampled for its intensity at discrete mass intervals (gray bars), shown for the $M + 8$ position of bromsulphthalein at a resolution of $R = 1.7 \times 10^5$. Pruning ($\beta = 0.2$, red dashed line) leads to a distorted envelope (red solid line), with the maximum distortion indicated in blue. Without normalization to measured intensities, envelopes are rescaled to range in [0,1].

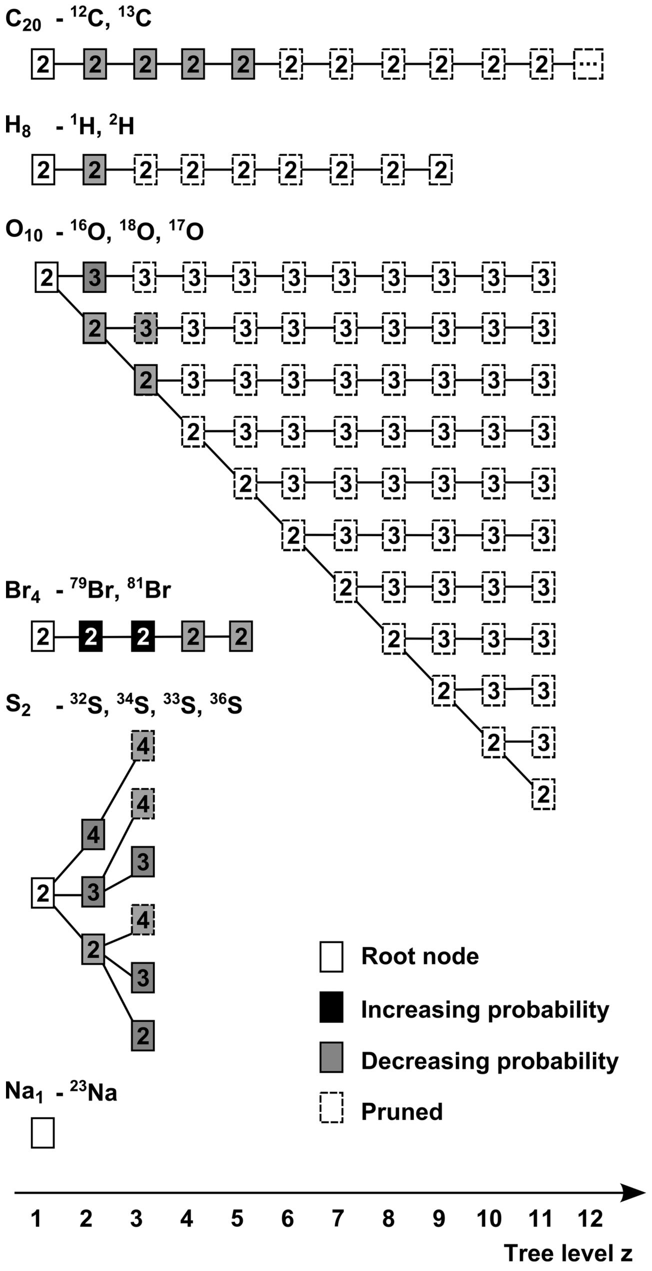

C_{20} - ^{12}C, ^{13}C
H_8 - ^{1}H, ^{2}H
O_{10} - ^{16}O, ^{18}O, ^{17}O
Br_4 - ^{79}Br, ^{81}Br
S_2 - ^{32}S, ^{34}S, ^{33}S, ^{36}S
Na_1 - ^{23}Na
□ Root node
■ Increasing probability
▨ Decreasing probability
⬚ Pruned
1
2
3
4
5
6
7
8
9
10
11
12
Tree level z

Figure 2. Transition trees for submolecules of the sodium adduct of bromsulphthalein at $\beta = 1 \times 10^{-5}$. Each node represents a subisotopologue and lists its transition index I, which points to the ordered subset of isotopes used for its transitions to child subisotopologues.

details on deriving isotopic envelopes can be found elsewhere.[8,12]

Transition Trees. The transitions of eqs 1 and 2 within each submolecule can be arranged into a treelike structure, henceforth termed a transition tree. These structures exclude redundant transitions and embrace all possible multicombinations of isotopes per element and thus the full set of subisotopologues. The transition trees of bromsulphthalein are shown in Figure 2. Each node in a transition tree represents a distinct subisotopologue. The root node at level $z = 1$ contains the subisotopologue composed of only the most abundant isotope $i = 1$ from the ordered set of isotopes. Each edge symbolizes a transition; that is, a replacement of one isotope $i = 1$ with another less abundant isotope $j > i$. As a consequence, a terminal node is reached at level $z = n_k + 1$ when no further isotopes $i = 1$ can be replaced; that is, $n_{k,1} = 0$.

In addition, each node contains a single index I, with $1 < I \leq q_k$. This so-called transition index determines all valid transitions from a parent at any level z to its child nodes at level $z + 1$ and must not be confused with a reference to a specific position of a node in the tree hierarchy. That is, all transitions with an isotope j under the restriction $I \leq j \leq q_k$ are feasible and must be conducted to grow the entire tree containing all possible subisotopologues. The transition index of a child node is in turn set to the j from which it was derived and is used in the very same manner for transitions to level $z + 2$. Overall, all nodes and their transition indices are recursively developed from the root node, which is initialized at $I = 2$ if more than one isotope exists for a submolecule. Moreover, trees collapse to a simple sequence for elements with $q_k = 2$ (e.g., carbon) or that contain none but the root node for $q_k = 1$ (e.g., sodium).

Pruning. The properties of transition trees allow for efficient pruning of low-probability subisotopologues, alias nodes, by use of either absolute or relative probability thresholds. For the latter case, probabilities are regarded as insignificant if they range below a certain fraction $0 \leq \beta < 1$ of the highest probability $P_{\max}$ among the isotopologues of a molecule:

$$P_{\max} = \prod_{k=1}^{k_{\max}} P_{k,\max} \tag{6}$$

where $P_{k,\max}$ is the highest subisotopologue probability in each submolecule k. Consequently, the determination of all $P_{k,\max}$ comes prior to any pruning, unless $\beta = 0$ is used to skip pruning. Each tree is therefore grown to reach its most probable node in a first phase. To this end, only transitions of increasing probability are made, starting from the root node. As can be seen in eq 1, the ratio $P_{k,y}/P_{k,x}$ of a transition is driven by $n_{k,1}/n_{k,j}$, as $a_{k,j}/a_{k,1}$ is constant for any given isotope $j > 1$. Once $n_{k,1}/n_{k,j} < a_{k,1}/a_{k,j}$ is reached at a given transition, all subsequent transitions are monotonically decreasing in their probabilities for isotope j because $n_{k,1}$ can only decrease and $n_{k,j}$ can only increase with level z in a transition tree. A node-specific check of index I for reference to any transition of increasing probability with isotopes $I \leq j \leq q_k$ indicates if transitions to all child nodes must be made in this first phase. Otherwise, all transition paths of higher-level branching from this parent node are of decreasing probability; the concerned node is then made temporarily dormant during this first phase. Moreover, the ordering of isotopes by their decreasing natural abundance helps to phase out branches of solely decreasing probability at low tree levels. In small molecules, the root node is often the most abundant subisotopologue, when $n_{k,1}/n_{k,j} < a_{k,1}/a_{k,j}$ holds for all isotopes $j > 1$. A second phase then progresses with the dormant transitions of strictly decreasing probabilities after $P_{k,\max}$ has been established for each submolecule (gray nodes in Figure 2). While doing so, all transitions from nodes with probability $P_{k,x}$ fulfilling eq 7

$$P_{k,x}\frac{\prod_{k=1}^{k_{\max}} P_{k,\max}}{P_{k,\max}} < \beta P_{\max} \tag{7}$$

or, after division by $P_{\max}$, eq 8

$$P_{k,x}/P_{k,\max} < \beta \tag{8}$$

are halted and downstream nodes are prepruned, even if no leaf nodes have been reached yet. The fraction on the left-hand side of eq 7 scales $P_{k,x}$ in submolecule k to the maximum probability it can attain in the later Cartesian product by setting probabilities in all other submolecules to their maximum. With no further transitions remaining, subisotopologue nodes that do not fulfill eq 8 are finally postpruned in a third phase.

The procedure outlined above simplifies for absolute probability thresholds, as no $P_{k,\max}$ need to be searched for in a first phase. Instead, all transitions from nodes indexing at least one transition of increasing probability have to be made. Transitions from nodes with probabilities below the absolute threshold and indexing only decreasing probabilities can be omitted. With no further transitions remaining, any developed nodes below the absolute threshold are again postpruned. Overall, a large set of isotopologues can be pruned in most trees (Figure 2, dashed nodes), greatly reducing the computational burden.

Further pruning can be applied during a third phase of combining the subisotopologues from different submolecules via their Cartesian products. Computationally, the joint probabilities are calculated by multiplying each subisotopologue probability of the first submolecule with each of the second submolecule. In turn, the resulting products are each multiplied with the subisotopologue probabilities of the third submolecule. This iteration continues until the last remaining submolecule has been included. The analogous summation of subisotopologue masses across the submolecules leads to the isotopologue masses of the full molecule.[9] After each of the $l = \{1, ..., k_{\max} - 1\}$ iterations, all products x with probability $P_{l,x}$ either fulfilling eq 9

$$P_{l,x}/\prod_{k=1}^{l+1} P_{k,\max} < \beta \tag{9}$$

for relative thresholds or ranging below the absolute probability threshold can be discarded, together with their related masses.

Parameter Evaluation. The presented algorithm, henceforth called *enviPat*, was first used to elucidate differences between relative and absolute pruning thresholds on the distortion of isotopic envelopes. For this purpose, 5969 unique organic molecular formulas were randomly sampled in mass bins from the PubChem database[14] and their theoretical envelopes were simulated for a combination of four complementary resolution functions (Thermo Orbitrap Elite, R = 240 000 at m/z 400 and 60 000 at m/z 400; Thermo Orbitrap Velos Pro, R = 7500 at m/z 400; Waters G2 QToF, R = 25 000 at m/z 200) with eight different absolute and relative thresholds each. Every simulation was then compared to an approximation of the corresponding unpruned envelope, after rescaling of each envelope to its most intense signal; that is, to range in [0,1]. The distortion was then recorded as the maximum intensity difference between the two scaled envelopes over all discretization points. An illustration is given in Figure 1. This approach imitates common practice in which a pruned simulated envelope is compared to an unpruned measured envelope after the former was scaled to the latter by use of the most intense signal. The approximated unpruned envelope contained a minimum cumulative probability $\geq$0.9995 from using a very low absolute probability threshold of 1×10^{-20}. The mass discretization of all envelopes was dynamically adjusted for each formula and resolution to $^1/_{10}$ the width at half-maximum of a Gaussian peak shape function.

Performance Comparison. *enviPat* was compared to two previous approaches for isotope fine structure calculation, *Isotope Calculator*[13] and *ecipex*,[9] using the above-mentioned benchmark set of molecular formulas. *ecipex* has only recently been published and also yields isotopic fine structures but uses convolutions via multidimensional Fourier transforms to generate all subisotopologues for each submolecule. Similar to *enviPat*, *Isotope Calculator* utilizes transitions between subisotopologues, but it organizes them into mass states. Subisotopologues are finally combined to form the isotopologues of the full molecule in all these approaches. Memory usage was therefore quantified by the sum of these intermediate subisotopologues and is hence independent of the specificities of memory allocation in the various implementations. For *enviPat*, this amounts to the number of pruned subisotopologues established at the end of phase two when growing the transition trees, summed over all submolecules. With no pruning available at this early stage, the full count of subisotopologues must be reported for both *ecipex* and *Isotope Calculator*. This clearly states a lower bound of memory requirements for *ecipex*, which in fact needs $(n_k + 1)^{q_k - 1}$ intermediate terms to be summed over all k submolecules. For a runtime comparison, we selected *ecipex* because it has already been shown to outperform *Isotope Calculator* in this aspect.[9] Furthermore, both *ecipex* and *enviPat* are implemented in the R statistical environment,[15] making a comparison more consistent.[16] For time performance, each approach was averaged over 200 repetitions for each formula, including an initial garbage collection step, a timeout of 2 s for a single isotope pattern calculation, and an upper limit of 2×10^7 subisotopologues. Because *ecipex* does not allow for relative thresholds, an absolute threshold from the above parameter evaluation was selected for both approaches to calculate equal numbers of isotopologues for each formula. Calculations were run in R version 3.0.2, on a workstation with a 2.5 GHz Intel core i5−3210 M processor and 8 GB RAM under Ubuntu 14.04, 64 bit.

■ RESULTS AND DISCUSSION

The evaluation of isotopic fine structure calculations has to date been based on relatively small sets of carefully selected molecular formulas, with at most 10 formulas.[9] In contrast, our randomly sampled benchmark set comprises several thousand molecular formulas from 30 to 25 000 u, including large organometallic molecules with polyisotopic elements such as Zn, Ti, or Gd. The size and complexity of this benchmark set enhances both the representativeness of our findings and the detection of scaling issues below.

Parameter Evaluation. Pruning of isotopologues leads to changes in the simulated isotopic envelope of a molecule and thereby to deviations from its measured and hence unpruned envelope after the simulated envelope is rescaled to match the latter by its most intense signal. Figure 3 shows this distortion for the benchmark formulas and a range of threshold values covering several orders of magnitude. Listed as the maximum difference in scaled intensity and hence constituting a fractional quantity, the absolute intensity difference between simulated and measured envelopes at a given distortion increases with increasing signal intensity. In general, an approximately log−linear relationship can be observed between the median fractional distortion and the threshold values, for both relative and absolute thresholds. The variation in distortion at a given threshold value, however, differs between threshold types, pooled over the different resolutions and our extensive benchmark set of molecular formulas. Namely, the overall spread in distortion is smaller for relative instead of commonly applied absolute thresholds. This translates to a less skewed distortion on a linear scale; that is, a much lower tendency to generate outliers of imprecise isotopic envelopes for the single threshold choice to be made. In contrast to an absolute choice, a single relative threshold relates to widely differing cutoff probabilities $\beta P_{\max}$ among the different molecular formulas (data not shown). For molecules with a large spread in probability among isotopologues, this cutoff is often lower and hence permits a larger number of relevant isotopologues to be included. For example, at $\beta = 1 \times 10^{-5}$, bromsulphthalein requires the calculation of 601 isotopologues, with $P_{\max} = 0.265$. On the contrary, the much larger molecule bovine insulin, $C_{254}H_{377}N_{65}O_{75}S_6$, needs 5456 isotopologues to be calculated, using the same relative threshold for $P_{\max} = 0.113$.

Performance Comparison. We selected an absolute probability threshold of 1×10^{-9} for a performance comparison from the above parameter evaluation. The concomitant maximum distortion of 6.7×10^{-5} (median 1×10^{-7}, cf. Figure 3) lies far below the intensity uncertainties of available MS instrumentation.[17] In any case, the general findings below were not compromised by alternative threshold choices. In combination with the chosen time and memory constraints, computation failed for only 11 and 69 of the benchmark formulas for *enviPat* and *ecipex*, respectively. The total number of molecular isotopologues given by eq 5 stretches over 14 orders of magnitude for the remaining formulas and thus sheds light on the performance scaling with increasingly complex molecules.

First, the number of subisotopologues assembled in intermediate calculations increases nonlinearly with the total number of isotopologues in a molecule. In the unpruned case, the former number is approximately 3 times lower in orders of

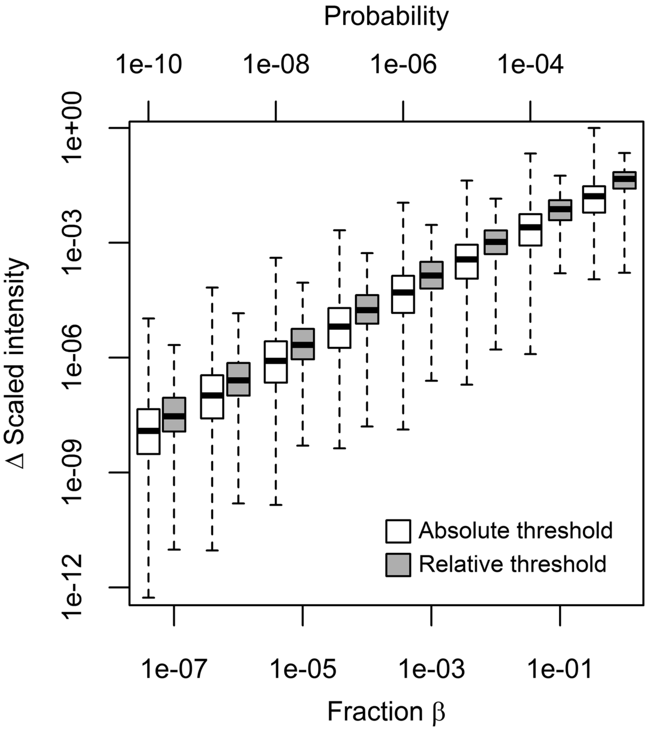


Figure 3. Difference in scaled intensity (distortion) of theoretical envelopes for different pruning thresholds, pooled over the four instrument resolutions and the benchmark set of molecular formulas (box plots indicate extremes, median, and lower and upper quartiles). Relative thresholds are given as fraction β of the most probable isotopologue, whereas absolute thresholds refer to cutoff probabilities. Note the log scale of all axes.

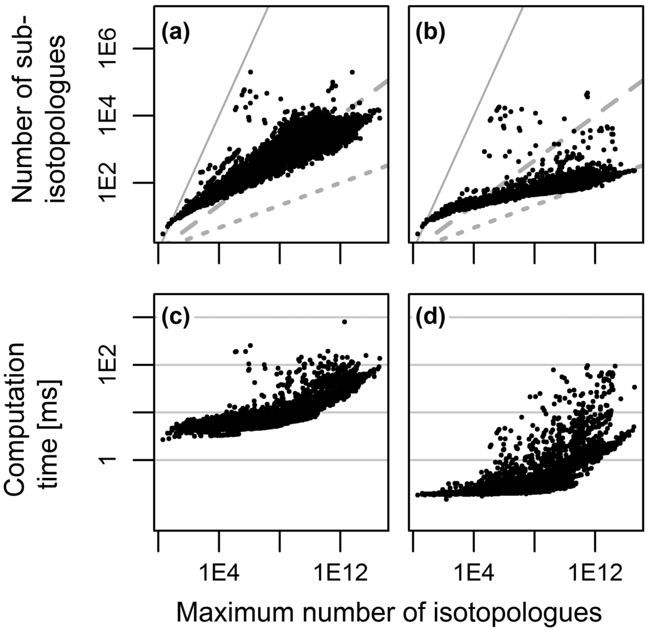


Figure 4. Number of (a) unpruned and (b) *enviPat*-pruned subisotopologues and computation time for (c) *ecipex* and (d) *enviPat*, plotted against the total isotopologue number of each molecule in the benchmark set (axes are log-scaled). Gray lines in the top plots indicate numbers of subisotopologues being equal (solid) or 3 (dashed) and 6 (dotted) times smaller in orders of magnitude than the total number of isotopologues.

magnitude relative to the latter, although this ratio increases for smaller molecules (Figure 4a). The variation in this relationship generally grows with more complex molecular formulas. In contrast, *enviPat* drastically reduces both the number of subisotopologues and its variation (Figure 4b). While the mentioned ratio is still similar for small molecules, it drops more strongly with rising molecular complexity than in the unpruned case. For the most complex molecules in our benchmark set, the number of subisotopologues can be as much as 6 times lower in orders of magnitude than the total number of isotopologues. Similarly, subisotopologue counts exceeded the numbers of molecular isotopologues above the given threshold in 37.5% of the cases (data not shown). This occurred for only 0.1% of cases for *enviPat*. Such favorable scaling is a result of the aggressive pruning that *enviPat* features. As demonstrated in Figure 2 for bromsulphthalein, branches containing large fractions of subisotopologues can be ignored, solely on the basis of probability, isotopic composition, and transition index of their parent node. In this way, subisotopologue differences of up to 2 orders in magnitude can be achieved in comparison to the unpruned approaches. This does not imply that all developed subisotopologues range above their threshold probability. Being directed by a single transition index per node, transitions of increasing probability can be accompanied by a set of decreasing and thus irrelevant ones, which have to be postpruned. Omission of tree branches will also be less efficient for molecules dominated by elements with several isotopes of similar natural abundance, explaining the few outliers of Figure 4b, containing, for example, Pt.

The overall numbers of subisotopologues will seldom exceed memory resources, even for the observed outliers. Unnecessary calculations and subsequent postpruning will rather affect computational runtimes, adding to methodological differences in deriving subisotopologues among the approaches. Indeed, *enviPat* outpaces *ecipex* for all benchmark molecular formulas, as shown in the bottom panels of Figure 4. Even for small molecules with similar numbers of subisotopologues in both approaches, *ecipex* computations take around 5 ms longer. One may argue that this constant overhead may result from implementation differences in, for example, molecular formula parsing, memory allocation or sorting instead or conceptual distinctions, giving *enviPat* an unfair advantage. However, this overhead would be small in relation to the logarithmic runtime differences between the two algorithms for more complex molecular formulas. Thus, even when the overhead is subtracted from the computation time of *ecipex*, *enviPat* still consumes around 1 order of magnitude less computation time for molecules with more than 1×10^6 isotopologues.

Transition Tree Properties. With *enviPat*, the accuracy of an isotopologue mass and probability is limited only by the computational precision and rounding errors. The latter increase with the number of calculation steps needed to calculate subisotopologues. For elements with more than two isotopes, for example, oxygen and sulfur of Figure 2, the branching within trees results in fewer steps to reach a node by transitions from the root node than a purely sequential updating, on average. But even for elements with two isotopes, for which trees simplify to sequences, probability differences between a node at level z and its parent at level $z - 1$ decrease monotonically with z. Any significant rounding errors will thereby accumulate in the pruned tails of these sequences long after the most probable subisotopologue was detected. A similar situation arises for isotopes with variable natural abundances (for instance, boron), where uncertainties in the probability of an isotopologue increase with the number of transitions. Here, finding the most dominant isotopologues within a small number of calculation steps is important to minimize uncertainties of the final envelope.

Several other properties of transition trees need to be stressed to direct future implementations. First, no information other than the state (i.e., mass, probability, isotopic composition) and transition index of a node is required to develop its transitions. The history of tree growth or the state of other nodes is irrelevant. This leaves a great potential for parallelization of the tree growth procedure ab initio (absolute threshold) or after the most abundant isotopologue has been detected (relative threshold). Similarly, parent nodes of undeveloped tree branches can be saved to hard disc for later evaluation under very stringent RAM conditions or for extremely large molecules. Second, transition trees can be built without division into submolecules. Instead, transition indices can reference the full set of isotopes over all elements in a molecule, excluding those of highest abundance per element, which are used to build the root node. Tree nodes then no longer represent subisotopologues of submolecules but directly the isotopologues of the full molecule. Herein, a different metric for sorting the referenced set of isotopes should be considered for quickly pruning branches of decreasing probability, for example, by use of the probability ratio P_y/P_x of the current transitions from node x to any of its children y. Alternatively, divisions can be made into submolecules containing more than one element, too. The above aspects can also be integrated to directly superimpose the peak shapes of molecular isotopologues after their transitions are developed, without a need to store them. This would drastically shrink the memory requirements for envelope calculations. Third, the order of an isotope subset referenced by transition indices does not need to be static. Rather, a set can be copied and reordered to meet the specificities of different branches, especially if transition trees for more than a single element may be built. Fourth, indices can be split. For example, given the index $I = 2$ at $q_k = 5$, only the transition $j = 2$ of, for example, increasing probability can be made, instead of the full referenced set 2, ..., 5. The concerned node is then set to $I = 3$ for later evaluation of transitions with decreasing probability. Some of these properties are further illustrated in Figure 5.

■ IMPLEMENTATION

The presented algorithm is freely available as R package *enviPat*. The package provides instrument-specific envelope and centroid calculation, batch processing, and molecular formula parsing for a variety of adducts commonly formed during electrospray ionization (ESI). User-defined inputs cover absolute and relative pruning thresholds, charge states, resolution-dependent envelope discretization, and enriched isotope tables, among others. To facilitate its usage, all *enviPat* functionalities and simultaneous comparison with measured data can be conveniently accessed from a Web-based user interface at www.envipat.eawag.ch.

■ CONCLUSION

Comparatively few strategies have been proposed to derive exact isotope fine structures for the simulation of mass spectrometric signals. While supporting the interpretation of highly resolved MS data, such strategies must be able to

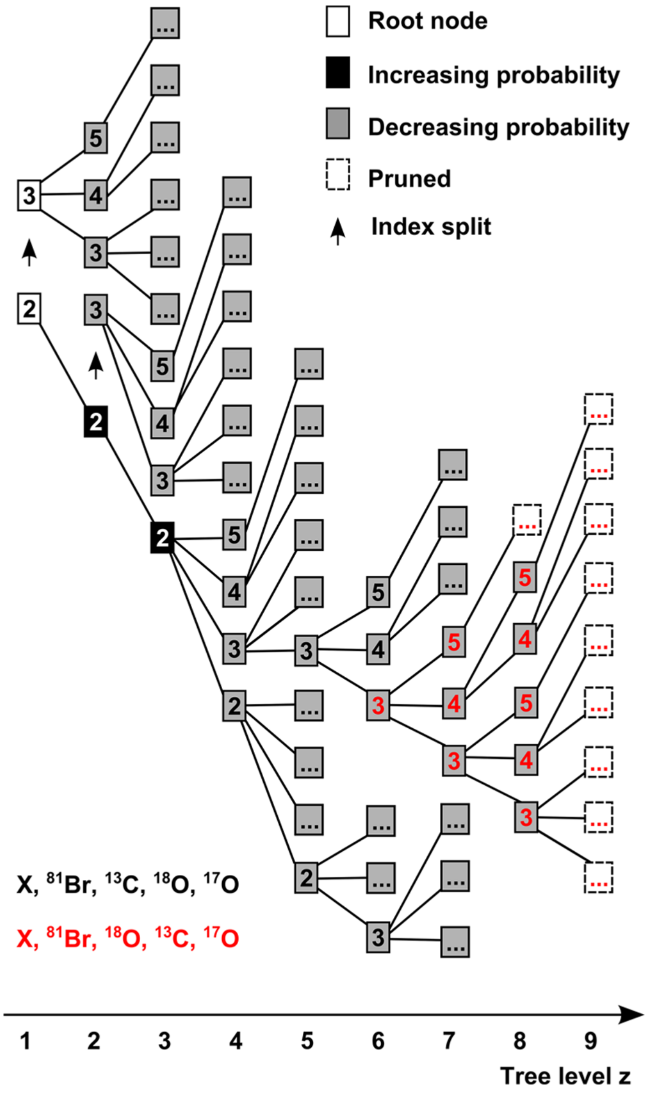


Figure 5. Transition tree for submolecule $C_{20}O_{10}Br_4$ of bromsulphthalein at $\beta = 1 \times 10^{-8}$. Here, transition indices of nodes refer to the joint set of all isotopes of the three elements, with X representing a single entry for the monoisotopic isotopes to be replaced. Non-monoisotopic isotopes are initially sorted by decreasing probability ratios P_y/P_x (black transition indices) and can be resorted for individual branches after several transitions (red indices). Moreover, arrows indicate a split of the transition index to separate transitions of increasing probability from decreasing ones.

efficiently prune large fractions of low-probability isotopologues. We therefore introduced a new hierarchy to calculate child from parent isotopologues by single replacements of one most abundant isotope by a less abundant one, guided by a node-specific transition index. As opposed to previous hierarchies, our treelike structure permits a precise pruning of large but low-probable isotopologue branches, solely based on the state of a single isotopologue at the root node of such branches. In addition, calculations can for the first time be directed to derive the most probable isotopologue in a first phase. Its probability can then be used for a relative instead of an absolute pruning threshold in a second phase of remaining calculations. On the one hand, this relative pruning adapts to the highly heterogeneous dispersion of isotopologue probabilities among diverse and often unknown analytes. On the other hand, pruning of substantial isotopologue branches leads to a computational performance that dwarfs the runtime required to convolute the resulting isotopologues to their measurable envelopes in practical applications. Hence, and despite several yet-unexploited properties of our so-called transition trees, this speed-limiting step should best be addressed next.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail martin.loos@alumni.ethz.ch.

Author Contributions

All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The work presented was partly financed by the Swiss Federal Office for the Environments (FOEN) and a mobility fellowship from the Swiss National Science Foundation (SNF).

■ REFERENCES

(1) Schymanski, E. L.; Jeon, J.; Gulde, R.; Fenner, K.; Ruff, M.; Singer, H. P.; Hollender, J. *Environ. Sci. Technol.* **2014**, *48*, 2097−2098.

(2) Roussis, S. G.; Proulx, R. *Anal. Chem.* **2003**, *75*, 1470−1482.

(3) Krauss, M.; Singer, H.; Hollender, J. *Anal. Bioanal. Chem.* **2010**, *397*, 943−951.

(4) Kenar, E.; Franken, H.; Forcisi, S.; Wörmann, K.; Häring, H. U.; Lehmann, R.; Schmitt-Kopplin, P.; Zell, A.; Kohlbacher, O. *Mol. Cell. Proteomics* **2013**, *13*, 348−359.

(5) Rockwood, A. L.; Kushnir, M. M.; Gordon, J. N. *J. Am. Soc. Mass Spectrom.* **2003**, *14*, 311−322.

(6) Valkenborg, D.; Mertens, I.; Lemière, F.; Witters, E.; Burzykowski, T. *Mass Spectrom. Rev.* **2012**, *31*, 96−109.

(7) Scheubert, K.; Hufsky, F.; Böcker, S. *J. Cheminf.* **2013**, *5*, 12.

(8) Werlen, R. C. *Rapid Commun. Mass Spectrom.* **1994**, *8*, 976−980.

(9) Ipsen, A. *Anal. Chem.* **2014**, *86*, 5316−5322.

(10) Snider, R. K. *J. Am. Soc. Mass Spectrom.* **2007**, *18*, 1511−1515.

(11) Yergey, J. *Int. J. Mass Spectrom. Ion Phys.* **1983**, *52*, 337−349.

(12) Li, L.; Kresh, J. A.; Karabacak, N. M.; Cobb, J. S.; Agar, J. N.; Hong, P. J. *J. Am. Soc. Mass Spectrom.* **2008**, *19*, 1867−1874.

(13) Li, L.; Karabacak, N. M.; Cobb, J. S.; Wang, Q.; Hong, P.; Agar, J. N. *Rapid Commun. Mass Spectrom.* **2010**, *24*, 2689−2696.

(14) Bolton, E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. *Annu. Rep. Comput. Chem.* **2008**, *4*, 217−241.

(15) R Core Team. R Project for Statistical Computing, 2014.

(16) Hu, H.; Dittwald, P.; Zaia, J.; Valkenborg, D. *Anal. Chem.* **2012**, *84*, 7052−7056.

(17) Guerrasio, R.; Haberhauer-Troyer, C.; Steiger, M.; Sauer, M.; Mattanovich, D.; Koellensperger, G.; Hann, S. *Anal. Bioanal. Chem.* **2013**, *405*, 5133−5146.