Sequence optimization as an alternative to de novo analysis of tandem mass spectrometry data

Alejandro Heredia-Langner1,∗, William R. Cannon2, Kenneth D. Jarman1 and Kristin H. Jarman1

1Statistical and Mathematical Sciences and 2Computational Biosciences, Pacific Northwest National Laboratory, Richland, WA 99352, USA

Received on December 7, 2002; revised on January 30, 2004; accepted on February 2, 2004
Advance Access publication April 15, 2004

ABSTRACT

Motivation: Peptide identification following tandem mass spectrometry (MS/MS) is usually achieved by searching for the best match between the mass spectrum of an unidentified peptide and model spectra generated from peptides in a sequence database. This methodology will be successful only if the peptide under investigation belongs to an available database. Our objective is to develop and test the performance of a heuristic optimization algorithm capable of dealing with some features commonly found in actual MS/MS spectra that tend to stop simpler deterministic solution approaches.

Results: We present the implementation of a Genetic Algorithm (GA) in the reconstruction of amino acid sequences using only spectral features, discuss some of the problems associated with this approach and compare its performance to a de novo sequencing method. The GA can potentially overcome some of the most problematic aspects associated with de novo analysis of real MS/MS data such as missing or unclearly defined peaks and may prove to be a valuable tool in the proteomics field. We assess the performance of our algorithm under conditions of perfect spectral information, in situations where key spectral features are missing, and using real MS/MS spectral data.

Contact: Alejandro.Heredia-Langner@pnl.gov

INTRODUCTION

The identification of peptides derived from complex mixtures of proteins is a prerequisite for several high-throughput proteomics technologies (Gavin et al., 2002; Ho et al., 2002; Link et al., 1999; Pasa-Tolic et al., 1999). Typically, protein mixtures are digested with trypsin and the resulting peptides are sequenced using tandem mass spectrometry (MS/MS). The sequence information provided by MS/MS analysis ideally consists of sequential mass-to-charge ratios (m/z, in units of Th) of the peptide as it fragments along the peptide backbone. The analytical challenge is how to use this sequence information in order to identify the unknown peptide.

In general, two approaches have been proposed for the solution of this problem. In the first, the MS/MS spectrum of an unknown peptide is compared with idealized spectra derived from genomic databases (Eriksson et al., 2000; Eng et al., 1994; Mann and Wilm, 1994). This method will fail to find a correct answer if the peptide sequence under investigation is unavailable in the search database due to differences in the genomes of the organism studied in the field and the one sequenced, frame shifts that occur during translation, alternative splicing or post-translational modifications.

The second approach attempts to find an amino acid sequence that would produce the spectrum at hand without referring to an archive of previously available peptide sequences. This de novo methodology uses only the peaks in the spectrum to deduce the sequence of amino acids that gave rise to it and is usually stated in a graph-theoretical framework (Bartels, 1990; Dancik et al., 1999; Taylor and Johnson, 1997; Hines et al., 1992; Chen et al., 2001).

To illustrate, consider the peptide LFSQVGK (Kinter and Sherman, 2000). A complete and perfect fragmentation of this peptide into singly charged b- and y-ions would produce peaks at the positions shown in Figure 1. The information contained in Figure 1 can be used to reconstruct the original peptide because the difference (in mass/charge, or m/z, units) between adjacent peaks of the same ion type corresponds to the mass of an amino acid residue in the original sequence for fragments with a charge of +1. If fragmentation occurs at every amino acid and every resulting fragment is detected as a singly charged ion, the problem of reconstructing the peptide using spectral information is greatly simplified and can be solved efficiently using dynamic programming methods (Dancik et al., 1999; Chen et al., 2001).

Unfortunately, experimental results are seldom this perfect and the researcher is usually confronted with spectra that...
Numerical values shown are approximations.

The implementation of our GA starts with a small initial population of amino acid sequences generated at random. These sequences are manipulated using mechanisms of recombination, selection and mutation until some criterion of convergence has been met.

**Development of the fitness function**

In most optimization problems, the objective or objectives, can usually be clearly stated as mathematical functions or some combination of rules to be followed or decisions to be made under appropriate circumstances. In the case of MS/MS spectra, all we know is that the end result should be a complete sequence whose mass and main spectral features match those in the experimental spectrum.

We could ask whether pursuing these objectives (total peptide mass and spectral peaks) individually may provide the algorithm with enough guidance to obtain a correct answer. An objective function that uses only total peptide mass as target will often result in a solution that matches the target mass well but has an incorrect sequence of amino acids. Direct comparison of features between spectra (the target and that from a potential solution) can also fail to produce the desired result because spectra from peptides that are structurally alike can be very dissimilar. The solution must ultimately match the target mass and target spectral features, but these cannot be the only guides during the search. The fitness function in this paper employs total peptide mass and target spectral features as part of a multiple-objective approach, but it is supplemented by a term that provides information about the similarity between two spectra. This similarity term counts the number of peaks between two spectra that could be matched by an appropriate translation. Efforts to derive similarity measures among mutated and modified peptides have been presented in the work of Pevzner et al. (2000, 2001). We develop a methodology to measure peptide similarities analogous to the cited references, but one that is adapted specifically for the GA.

Our objective function can be written as

$$\text{fitness} = w_1 \cdot \sum \text{match peaks} - w_2 \cdot \sum \text{non match peaks} + \frac{w_3}{1 + |wt - target|} + w_4 \cdot \text{similarity index},$$

where $w_1, w_2, w_3$ and $w_4$ are empirical constants whose magnitude can be adjusted to alter the relative importance of every term in the fitness function; target is the target parent peptide mass and $wt$ is the total mass of the peptide under evaluation. The first two terms in the objective function measure how well the target spectrum matches the one produced by a candidate peptide. These terms reward features of a potential solution that match peaks (within some tolerance) in the target spectrum and penalize unmatched peaks. In this paper, we consider complete fragmentation of the peptide representing the potential answer into b- and y-ion fragments only, but other ion types can also be considered. The third term penalizes deviations from target mass.

The first three terms in the objective function are traditional, and relatively simple, measures of similarity between spectra.

**SYSTEMS AND METHODS**

**Specific approach for MS/MS data**

The implementation of our GA starts with a small initial population of amino acid sequences generated at random. These sequences are manipulated using mechanisms of recombination, selection and mutation until some criterion of convergence has been met.
However, these three terms alone can in many cases fail to identify peptide chains that are nearly identical to the one that produced the target spectrum and are, for this reason, of limited help to the GA. The spectra of two peptides that differ by a single amino acid can potentially result in more non-matching spectral features (including total mass) than spectra produced by amino acid chains whose contents differ in more dramatic ways. This means that the first three terms in the objective function can drastically drop a solution from a high fitness value by a change of a single amino acid in the correct sequence. From this, it follows that we need to supply our fitness function with a way to discriminate between peptides that are at least partially good matches for our target from those that are not.

To this end, we introduce a fourth term in the fitness function in an attempt to derive a measure of overall similarity between two spectra. Consider an experimental spectrum whose \( m/z \) values can be described as a set of \( m \) peaks, \( S = \{s_1, s_2, s_3, \ldots, s_m\} \) (possibly consisting of more than \( b \)- and \( y \)-series ions), and the simulated spectrum of a potential solution, \( P = \{p_1, p_2, p_3, \ldots, p_n\} \) (as used in this paper, consisting of \( b \)- and \( y \)-series ions only), as a set of \( n \) peaks. Computing the differences between \( m/z \) values of every peak in \( S \) and every peak \( P \) in results in an \( m \) by \( n \) matrix of differences, \( D = \{d_{ij} = (s_i - p_j)\} \), \( 1 \leq i \leq m, \ 1 \leq j \leq n \) whose entries can be inspected to find peaks in \( P \) that, if translated, would match those in \( S \). If every entry in \( D \) has a distinct numerical value, it is not possible to simultaneously match more than one peak between \( P \) and \( S \) when adding a single real number to all the \( m/z \) entries of either spectrum. On the other hand, if multiple entries in \( D \) have the same numerical value (within a certain tolerance), they represent peaks in \( P \) that—either in their original position or after an appropriate translation—can be made to match peaks in \( S \) simultaneously. To illustrate this concept with a trivial example, consider what would happen if the numerical entries for \( S \) and \( P \) were obtained using the same peptide chain. In that case, the matrix \( D \) would possess a large number of zeros \( (m, \text{if the size of both} \ S \text{and} \ P \text{is} \ 1 \text{by} \ m) \). If the peptide chains represented by \( S \) and \( P \) differ by a single amino acid, the \( D \) matrix formed by this new pair of sequences would probably have few zeros (depending on where the substitution has occurred) and a relatively large number of entries with a numerical value equal to the difference between the masses of amino acids involved in the dissimilarity between chains. The multiplicity of repeated entries in \( D \) can be used as an indication of the similarity level between \( S \) and \( P \).

Pevzner et al. (2000) considered cases where the shifts needed to match peaks between two spectra could be attributed to the substitution of one or two amino acid residues in the original peptide chain. Their procedure for spectral alignment is based on a dynamic programming algorithm where both spectral peaks and the masses of amino acids are approximated by integers. Their work considers only ions in the \( b \)-series since simultaneous use of \( b \)- and \( y \)-series ions (or other types of ions) can make the dynamic programming algorithm converge to infeasible solutions. In our case, we are not interested in finding a particular amino acid substitution that can be used to explain all the differences between two spectra. Instead, our aim is to use the number of repeated entries in \( D \) to help us assess the relative fitness of potential solutions to our problem. The last additive term in the objective function incorporates information regarding whether two or more peaks in the simulated spectrum of a potential solution can be simultaneously matched to those in the target spectrum by an appropriate translation. The similarity index in our objective function is defined as the number of non-distinct entries in the matrix \( D \) (within a given tolerance) divided by the total number of elements in \( D \). This last term in the fitness function increases in value only if at least two peaks can be simultaneously matched by a translation.

As a simple test of the potential usefulness of the similarity term in the objective function, a histogram of the number of peaks that can be matched by translation (as measured by the number of non-distinct entries in the matrix \( D \)) for 10 000 independently and randomly generated amino acid sequences (with length between 7 and 10 amino acids each) and the target spectrum in Figure 1 is shown in Figure 2. In this figure, we also show histograms of the number of non-distinct entries in the matrix \( D \) found when comparing the idealized spectrum of Figure 1 and the simulated spectra produced after one, two and three amino acid substitutions in the sequence LFSQVKG.

A large majority of the randomly generated sequences have a number of non-distinct entries in \( D \) of 30 or less. This is also true for the histogram obtained when three amino acids are simultaneously substituted in the LFSQVKG sequence. The other two cases, one and two substitutions in the correct sequence produce a distribution of non-distinct entries in \( D \) with higher mean. Considering that a change of three amino acids in LFSQVKG corresponds to alterations in over 40% of its contents, the differences in these histograms indicates that significant damage to the original peptide sequence must be done before the distribution of non-distinct entries in \( D \) resembles that seen with randomly generated amino acid chains. Ideally, we would prefer that the distributions shown in Figure 2 would not overlap at all. As it is, this figure suggests that the fourth term in the fitness function can be helpful in distinguishing sequences that are similar to the one that produced the target spectrum from those that are not.

To test the effectiveness of our approach, we employ a series of runs with three different scenarios: (1) perfect spectral information; (2) the case in which key peaks are missing; and (3) a series of real MS/MS spectra. Perfect information. The full spectrum in Figure 1 is used as the target (using \( m/z \) values only, i.e. no intensity information is employed).
Sequence optimization of tandem mass spectrometry data

Fig. 2. Distribution of the number of peaks (shown in the x-axis) that can be matched by a translation between the target spectrum in Figure 1 and, clockwise from the top left: (a) 10,000 randomly generated sequences (length 7–10), (b) one, (c) two and (d) three amino acid substitutions in the sequence LFSQVGK.

Missing peaks. All combinations of one and two peaks in the spectrum shown in Figure 1 are deleted and the algorithm executed as above.

Real MS/MS spectra. In the last example, we use real spectra and compare our results with those obtained using the same input data and a de novo sequencing algorithm. The spectra were obtained from both literature and in-house MS/MS analysis at the Pacific Northwest National Laboratory in the laboratory of Richard Smith. The latter spectra were analyzed using a current best practice for identifying peptides from LCQ MS/MS data (Wolters et al., 2001). Each peptide was analyzed multiple times on multiple days with a Finnigan ion-trap mass spectrometer and at least one spectrum for each peptide resulted in Sequest scores exceeding 2. In addition, the mass of each peptide parent ion was examined as to whether it confirmed to within one part-per-million of the theoretical mass for that peptide by the use of an 11.5 T ion-cyclotron resonance mass spectrometer and a 15% elution time tolerance (Smith et al., 2002).

To assess the quality of the results obtained with our procedure, we have used Lutefisk (Taylor and Johnson, 1997, 2001), a de novo sequencing program, for comparison purposes. Lutefisk, one of many de novo algorithms available, was chosen in this paper due to its ease of use and access and because it is representative of the way in which many other de novo methodologies work.

In general, Lutefisk attempts to reconstruct complete peptides by extending candidate sequences, one amino acid at a time, in a way that maximizes a score function. Only sequences that match the total peptide mass (within certain tolerance) are stored. A maximum of five top-scoring
sequences are usually reported, although a smaller number is common. Whenever the program finds gaps in the spectrum that cannot be identified with certainty, Lutefisk inserts the mass of the gap in square brackets into the final sequence. A complete description that does full justice to the way in which Lutefisk functions is beyond the scope of this paper but the interested reader will be able to find resources in the published work (Taylor and Johnson, 1997, 2001).

RESULTS AND DISCUSSION

Example 1: testing the performance of the GA under conditions of perfect information

The target for our first example is the complete, perfect spectrum used in Figure 1, consisting of the set of peak locations given in Table 1 (which shows the target spectra for this and the rest of the examples in this paper). The target mass of the full peptide is 778.91 Da, which is the \( m/z \) value of the precursor peptide (389.46) multiplied by a charge of 2.

The values of the weights in the fitness function for the first example were (empirically) set to \( w_1 = 1 \), \( w_2 = 1 \), \( w_3 = 1 \) and \( w_4 = 20/|D| \), where \( |D| \) represents the number of elements in the matrix \( D \). A matching peak was counted if a simulated peak was within 0.01 Da of a feature in the target spectrum. The strict tolerance for counting matching spectral features was used because we know that there is no error in the measurements of the target spectrum. Out of 1000 independently started runs with perfect information, the correct sequence was found at the end of 384 of them. These 384 correct sequences also had the highest score among all 1000 independently started runs with perfect information, the correct sequence was found at the end of 384 of them. These 384 results. The tolerance values used for the GA we propose is not automatically deterred by missing

Example 2: missing features in an otherwise perfect spectrum

Our second scenario corresponds to a situation commonly found in practice, and a hurdle for many de novo sequencing algorithms.

One and two peaks at a time were deleted from the target spectrum in Example 1 and the resulting information entered into the GA. A total of 10 independently started runs were made after every deletion. The rest of the parameters for the algorithm were kept as those in Example 1.

In all the answers found after every group of runs, the fitness values of the answers were highly correlated with their quality. For example, when the 114.16 entry was deleted from the original spectrum, the sequence LFSQVGTK was found four times (with the highest fitness among all 10 reported answers) followed by LFSQVGTK (three times) and three other, less fit answers. Runs where a different peak was missing from the target spectrum yielded very similar results. A summary of the number of times that the correct sequence was found in every set of 10 runs when one and two peaks in the target spectrum were missing is shown in Table 2. The numbers shown in Table 2 do not convey a sense of the quality of the answers produced by the GA since counting only the number of perfect solutions dismisses the fact that all the peptides reported as answers could be partially matched to relatively large portions of the available data. The peptides obtained in these runs are structurally very similar to those shown in the previous paragraph.

It is apparent from the results in Table 2 that the version of the GA we propose is not automatically deterred by missing

---

Table 1. Spectra used in each example discussed in the text along with the mass of the parent ion

<table>
<thead>
<tr>
<th>Spectrum/example</th>
<th>Parent mass</th>
<th>Fragment ion masses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 2</td>
<td>778.91</td>
<td>114.16, 147.18, 204.23, 261.34, 303.36, 348.42, 431.49, 476.55, 518.57, 575.68, 632.73, 665.75</td>
</tr>
<tr>
<td>3</td>
<td>965.10</td>
<td>147.2, 245.2, 260.3, 358.3, 423.5, 471.5, 494.6, 542.6, 607.1, 705.8, 720.9, 819.8, 965.1</td>
</tr>
<tr>
<td>4</td>
<td>1738.00</td>
<td>305.5, 324.5, 509.5, 622.5, 660.3, 676, 693.5, 776.5, 788, 821.5, 843, 860, 879, 917, 1050, 1028, 1045, 1059, 1078, 1116, 1229, 1391.0, 1415.0, 1492</td>
</tr>
<tr>
<td>5</td>
<td>1825.90</td>
<td>375.1, 474.3, 520.3, 529.2, 573.5, 625.2, 679.1, 680.0, 681.4, 736.7, 745.4, 756.4, 819.7, 823.5, 832.4, 844.3, 848.2, 947.4, 948.4, 992.1, 1062.1, 1079.4, 1110.5, 1111.5, 1182.6, 1183.6, 1228.9, 1295.6, 1350.0, 1396.6, 1409.8, 1410.8, 1416.9, 1432.5, 1450.0, 1451.6, 1523.8</td>
</tr>
</tbody>
</table>

---

A.Heredia-Langner et al.
Sequence optimization of tandem mass spectrometry data

Table 2. Summary of results where the target spectrum supplied to the GA contained one or two missing values

<table>
<thead>
<tr>
<th>Peak missing</th>
<th>114.16</th>
<th>147.18</th>
<th>204.23</th>
<th>261.34</th>
<th>303.36</th>
<th>348.42</th>
<th>431.49</th>
<th>476.55</th>
<th>518.57</th>
<th>575.68</th>
<th>632.73</th>
<th>665.75</th>
</tr>
</thead>
<tbody>
<tr>
<td>114.16</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>147.18</td>
<td>5</td>
<td>3</td>
<td>7</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>204.23</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>261.34</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>6</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>303.36</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>348.42</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>431.49</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>476.55</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>518.57</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>575.68</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>632.73</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>665.75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Entries in the table are the number of times the correct sequence was found in 10 independently started runs of the GA.

Table 3. Summary of results where the target spectrum supplied to Lutefisk contained one or two missing values

<table>
<thead>
<tr>
<th>Peak missing</th>
<th>114.16</th>
<th>147.18</th>
<th>204.23</th>
<th>261.34</th>
<th>303.36</th>
<th>348.42</th>
<th>431.49</th>
<th>476.55</th>
<th>518.57</th>
<th>575.68</th>
<th>632.73</th>
<th>665.75</th>
</tr>
</thead>
<tbody>
<tr>
<td>114.16</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>147.18</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>204.23</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>261.34</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>303.36</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>348.42</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>431.49</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>476.55</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>518.57</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>575.68</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>632.73</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>665.75</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>

‘Y’ represents a run where the correct sequence was found, otherwise, the entry is ‘N’.

or incorrect information although, naturally, the quality of the solutions obtained will depend on the completeness and accuracy of the input data and on the parameters chosen for the algorithm.

When using Lutefisk with the datasets missing one and two peaks, we obtained the results shown in Table 3.

Since, in many cases, Lutefisk’s answer consists of a single sequence, results in Table 3 indicate only if the correct peptide chain was found (i.e. if the correct peptide was among the final solutions reported, regardless of rank). Lutefisk reports LFSKVGK as the best sequence but, for reasons already mentioned, we considered this to be the correct answer. As with the GA, results in Table 3 are somewhat simplified since in most cases the chain (or chains) reported by Lutefisk matched portions of the target sequence well.

In a majority of cases in Table 3, Lutefisk has not been able to find the correct peptide. This is not completely surprising because Lutefisk cannot bypass significant gaps present in the input data.

We now shift our attention to examples using real spectra. These examples will show how our algorithm performs on data with varying degrees of completeness, levels of noise and redundant or superfluous information.

Example 3: spectrum from the literature

Snyder (2000) presents, in one of many examples, a list of observed ions in the ESI-tandem mass spectrum of the T15 tryptic fragment of equine cytochrome c, corresponding to the amino acid sequence EDLIA YLK. The locations of the observed fragment ions are given in Table 1. This list of m/z data is an incomplete set of spectral features and was used as input to our program, along with the target mass (M + H) of 965.1 also reported by Snyder (2000). The weights in our objective function were set at $w_1 = 1$, $w_2 = 0.1$, $w_3 = 10$, $w_4 = 10/|D|$ and the tolerance for counting a
match between the target and simulated spectrum was set at 0.5 Da.

Of 40 independently started runs, the correct sequence was found on two occasions and was the highest scoring peptide among all 40 answers (final score of 22.9945). Naturally, some ambiguity remains in the final answers since our program was not able to distinguish between Ile and Leu. The top 10 distinct sequences found (in descending order of fitness) are EDLLAYLK, MLLLAYLK, MKYALLLL, MKYALPNK, MLLHSFLQ, MLLAYLMLAYL, VAY-HAYLQ, MLLAYLMLAYQL and MLLLAYLMLAYL.

Note that most of the reported peptides share similar subsequences and that the stiff penalty for deviations from target weight has helped to maintain the length of most of the reported peptides within reasonable limits. The number of peaks in the target spectrum matched by each of the top 10 answers is 11, 11, 9, 9, 8, 12, 8, 10 and 10, respectively.

The same \( m/z \) values were used as input for Lutefisk. Once again, tolerance levels of 1.0 Da for total peptide mass and 0.75 Da for fragment error produced the best results among our trials. A single sequence, FPLLAYLK was reported at the end of the Lutefisk run. This sequence is very close to the target one and the amino acid combination FP at the beginning of the peptide has almost the same mass as ED (the correct first two amino acids) or ML (reported in a majority of the GA solutions).

**Examples 4 and 5: spectra analyzed at PNNL**

**Example 4.** Peptides were derived from *Deinococcus radiodurans* by tryptic digestion and mass analyzed in the laboratory of Richard Smith at the Pacific Northwest National Laboratory. The \( m/z \) values reported in this, and all subsequent examples, were obtained after using a simple, statistically based procedure to eliminate some spectral features, possibly caused by noise, from actual MS/MS spectra (Jarman et al., 2003). The \( m/z \) values for this example are given in Table 1. The target mass for the parent peptide \( (M+H) \) is 1738 Da.

For this problem, we assigned weight values of \( w_1 = 1, w_2 = 0.1, w_3 = 10, w_4 = 5/|D| \) and a tolerance of 0.75 Da for counting matching peaks. These weights slightly reduce the relative importance of the similarity term in the fitness function as compared with the value used in Example 3. The rest of the GA parameters are the same as those used in Example 1.

Sequest’s identified solution is ANHWLAQGAAQTDTAR, with a total mass \( (M+H) \) of 1737.86 Da, \( \Delta Cn \) of 0.0000 and cross-correlation equal to 4.3391. As with the remainder of the examples, the mass of the precursor ion was confirmed to be correct to within one-part-per-million of the theoretical mass for that peptide by the use of an 11.5 T ion-cyclotron resonance mass spectrometer (Smith et al., 2002). Consideration of these two sources of evidence makes for a very strong case that the correct peptide has been identified during the database search.

The best 15 solutions out of 150 independently started runs of the GA are:

\[
\begin{align*}
Y & S & P & R & V & G & A & G & A & G & A & L & Q & G & C & T & T \\
H & E & G & A & N & L & A & Q & G & A & Q & G & Q & Q & L & Q & S \\
& Y & C & G & G & Q & L & A & Q & G & G & F & G & F & C & T & N & N \\
& H & G & K & W & L & A & Q & G & A & Q & G & A & Q & A & Q & A & D \\
& H & D & Q & Q & L & A & Q & G & A & Q & G & Q & G & D & R \\
& H & W & K & G & L & A & K & G & F & F & L & F & K & G & T \\
& H & E & G & A & Q & V & A & Q & G & Q & A & V & Q & K & G & T \\
& H & G & K & Q & G & N & A & Q & G & A & Q & G & L & G & A & K & S \\
& N & H & P & C & G & L & A & G & A & G & A & K & Q & Q & Q & E \\
\end{align*}
\]

with total mass \( (M+H) \) ranging from 1737.36 to 1738.09 Da. The middle portions of a majority of the sequences (Sequest’s and those found by the GA) are very similar. This is not surprising since real spectra usually possess more information available for the reconstruction of these middle peptide fragments than those at either end of the amino acid chain. The researcher should be aware that, under these circumstances, the amino acids present at either end of a peptide created by this sequence optimization approach or de novo algorithms can probably be arranged in different ways and still match the available target features equally well.

Lutefisk’s answers to this same set of \( m/z \) values (using tolerance levels of 1.0 Da for total peptide mass and 0.75 Da for fragment error) are: [185.0]HWLG[200.0]AE[312.2]TAR, [185.0]HWLG[200.0]AE[312.2]TVK, [304.0][317.1]AKEYE [312.2]TAR, [304.0][317.1]AKEYE[312.2]TVK and P[225.2]WLG[238.1]KA[275.0]TAR.

Peptides reported by Lutefisk match a maximum of six non-sequential amino acids in the database answer whereas several of the GA sequences match six and, in one case, seven consecutive amino acids present in the database answer. Nonetheless, it is interesting to note that several of Lutefisk’s answers report the (assumed) correct last three amino acids in the sequence, TAR, which the GA has not been able to detect. The parameter file we have employed for Lutefisk specifies tryptic digestion and therefore the de novo program forces an arginine or lysine residue at the end of the peptides while the GA does not (although it could be easily modified to do so).

**Example 5.** This example deals with a spectrum that has redundant peak information (in the form of peak positions that are slightly more than one \( m/z \) unit apart) and a relatively large number of data points. The 10 best sequences...
found out of a run with 100 independent starts (using the same objective function parameters as in the previous example and a tolerance of 1 Da), are YGMVVFVGSCLDLGPK, DDLGAGGVGRSCGLGQPMP, DWLNNAYNLFVVLGSR, TWQNNLCCYAAAPDPE, WTSVAPSMEGGFLLEG, DQMPGSEVCPVGCNGQGK, KDGNAVGSDDRSGGGNAN, FLLYVYLVMLCDCYK, MENTCRLGSGCLTDYM and PDDSGGLCSDYDSCDSTS.

Compare these sequences to the one reported by Sequest (DANNNAYDSDGVLRS) as the highest score found by database search. Sequest’s answer has a total mass \( M + H \) of 1824.8 Da. The GA answers have total mass ranging from 1823.2 to 1825.1 Da. The third answer reported by the GA matches the one found in the database search relatively well (10 amino acids in the GA solution can be matched to corresponding amino acids in the database answer at roughly the same positions).

The rest of the sequences found do not have much in common with either the database answer or each other and we decided to use this opportunity to investigate the behavior of our algorithm when provided with an initial population biased toward a specific portion of the problem space. To do this, we inserted the three highest scoring sequences found by the GA in the initial population of a new set of runs. The best five sequences from 100 independently started GA runs with this biased initial population are WNNNAYNYPYVNNQ, DWLNNAYNYPYVVNNQ, DWLNNAYNLFVVGAGGGG, DWLNNAYNLFVVGAG and DWLNNAYNLFVVGAGNN.

The introduction of highly fit individuals into the population has had the effect of narrowing the scope of the search and has produced nearly identical sequences after every run. The exercise of biasing the initial population is important because we plan to introduce portions of amino acid chains.

FUTURE WORK

There are several areas where improvements and further developments are possible. Our priority is combining the GA with a probability-based fitness function that allows scoring peptides on a likelihood scale (Jarman et al., 2003). Development of a meaningful and useful scoring function is vital if a performance comparison with de novo techniques, or other peptide identification algorithms, is to take place.

CONCLUSIONS

We have used an optimization methodology that attempts to gradually construct peptide sequences that match the target spectrum optimally. The GA implemented here is not immediately deterred by incomplete spectra, peaks produced by unusually occurring peptide fragments or background noise.

COMPUTATIONAL EFFICIENCY

As we have implemented it, our version of the GA creates 350 individuals in each of 150 generations, for a total of 52,500 evaluations of the objective function before reporting an answer. Considering that at least 10 (for the simpler examples) and at most 150 independently started runs have been used to find a series of potential sequences from which a final peptide can be selected, our algorithm goes through a number of sequences ranging from a little over half a million to close to 8 million to build a small set of potentially good amino acid chains.

For the examples we have presented here, strings with at least 10 and at most 20 spaces were considered. Our set of building blocks consists of 19 different amino acids (we cannot distinguish between Leu and Ile) and a blank character. The number of distinct candidate peptide chains available under these conditions ranges from \( 20^{10} \sim 1.024 \times 10^{13} \) for the strings with 10 spaces to \( 20^{20} \sim 1.049 \times 10^{26} \) for the examples with initial chains of up to 20 spaces. This means that our algorithm has explored at most \( 5.13 \times 10^{-6} \) of the available space (for the 10 amino acid chain and 10 solutions) and at most \( 7.51 \times 10^{-18} \) of the space (for the 20 amino acid chain and 150 generated solutions) before reporting an answer.

One distinct advantage of the GA is that the user has the ability to choose convergence criteria prior to an actual run and therefore determine the computational effort employed in obtaining a solution. The GA has proved that it can remain practically useful for problems which grow exponentially with the number of decision variables in the areas of reliability engineering (Coit and Smith, 1996) and experimental design (Heredia-Langner et al., 2003). This means that the algorithm has been shown to be capable of finding good answers in problems with very large spaces without using (a population that increases exponentially with the size of the problem).
Another development presented is the characterization of the problem space as an entity amenable of manipulation through the construction of an appropriate objective function. Traditional solutions to the peptide-sequencing problem work through a strictly defined objective function and often lack flexibility to deal with the realities of MS/MS spectra, requiring an exponential increase in computing effort as the complexity of the problem grows. Redefining the way in which the objective is presented is a first step in bringing new solution methodologies that can be helpful when dealing with this difficult problem.

ACKNOWLEDGEMENTS

The authors are grateful for the comments received from an anonymous referee which greatly improved the clarity and quality of this paper. This research was funded by the Office of Biological and Environmental Research in the US Department of Energy under contract 41966A and performed in the William R. Wiley Environmental Molecular Sciences Laboratory at the Pacific Northwest National Laboratory. The EMSL is funded by the Office of Biological and Environmental Research in the US Department of Energy. PNNL is operated by Battelle for the US Department of Energy under Contract DE-AC06-1830.

REFERENCES