Mutation-Tolerant Protein Identification by Mass-Spectrometry

Pavel A. Pevzner * Vlado Dančík† Chris L. Tang‡

Abstract

Database search in tandem mass spectrometry is a powerful tool for protein identification. High-throughput spectral acquisition raises the problem of dealing with genetic variation and peptide modifications within a population of related proteins. A method that cross-correlates and clusters related spectra in large collections of uncharacterized spectra (i.e from normal and diseased individuals) would be extremely valuable in functional proteomics. This problem is far from being simple since very similar peptides may have very different spectra. We introduce a new notion of spectral similarity that allows one to identify related spectra even if the corresponding peptides have multiple modifications/mutations. Based on this notion we developed a new algorithm for mutation-tolerant database search as well as a method for cross-correlating related uncharacterized spectra. The paper describes this new approach and its applications in functional proteomics.

1 Introduction

Tandem mass spectrometry (MS/MS) is a widespread method for identifying and analyzing proteins. At the first stage of MS/MS parent peptides (formed by enzymatic cleavages at specific sites along the backbone of a protein) are introduced to a mass spectrometer and ionized so that their mass/charge ratios may be measured At the second stage of MS/MS individual peptide ions may be selectively isolated and further fragmented to provide information about the mass/charge ratios of resulting fragment ions (tandem spectrum of a parent peptide). This simple description hides the details of the significant technical barriers in MS/MS that have been overcome only recently. As a result of these developments tandem mass-spectrometry is becoming a method of choice in many areas of proteomics.

*Departments of Mathematics, Computer Science, and Molecular Biology, University of Southern California, Los Angeles, CA 90089

†Millennium Pharmaceuticals, 640 Memorial Dr , Cambridge, MA 02139

‡Millennium Pharmaceuticals, 640 Memorial Dr , Cambridge, MA 02139

RECOMB 2000 Tokyo Japan USA

Our work was motivated by the following problem: given a large collection of uninterpreted experimental spectra find out which spectra in the collection correspond to similar peptides, i.e. peptides that differ by a small number of mutations/modifications. The algorithm that cross-correlates and clusters related spectra in large collections of uncharacterized spectra would be extremely valuable in functional proteomics. To the best of our knowledge, no such algorithm was described yet.

This functional proteomics problem is related to database search in mass-spectrometry that has been very successful in identification of already known proteins. Experimental spectrum can be compared with theoretical spectra for each peptide in a database and the peptide from the database with the best fit usually provides the sequence of the experimental peptide (Mann and Wilm, 1994 [13], Eng et al., 1994 [7], Clauser et al.,1996 [4], Taylor and Johnson, 1997 [16], Fenyo et al., 1998 [8]). In particular, SEQUEST ([7]) has been used to identify proteins from class II MHC complex while MS-Tag (Clauser et al., 1999 [3]) successfully identified proteins important to placental hypoxic response relevant to modeling the effects of preeclampsia However, in light of the dynamic nature of samples introduced to a mass spectrometer and potential multiple mutations/modifications, the reliability of the database search methods that rely on precise or almost precise matches may be called into question. *De novo* algorithms that attempt to interpret tandem mass spectra in the absence of a database (Johnson and Biemann, 1989 [12], Bartels, 1990 [1], Fernandez-de-Cosio et al., 1995, [9], Taylor and Johnson, 1997 [16], Dancik et al., [5]) are invaluable for identification of unknown proteins, but they are most useful when working with high quality spectra. Usually, the peptide must have good fragmentation and should not contain modified amino acids (Dancik et al., 1999 [5]).

Since proteins are parts of complex systems of cellular signalling and metabolic regulation, they are subject to almost uncountable number of biological modifications (such as phosphorylation and glycosylation) and genetic variation (Gooley and Packer, 1997 [11]) For example, at least 1000 kinases exist in the human genome, indicating that phosporylation is a common mechanism for signal transmission and enzyme activation. Almost all protein sequences are post-translationally modified and as many as 200 types of covalent modifications of amino acid residues are known. Since currently post-translational modifications cannot be inferred from DNA sequences, finding them will remain an open problem even after the human genome is completed. It also raises a challenging computational problem for post-

genomic era: given a very large collection of spectra representing the human proteome, find out which of 200 types of modifications are present in each human gene. Below we describe a computational approach to this problem based on a new mutation/modification-tolerant database search.

Starting from the classical Biemann and Scoble, 1987 [2] paper there were a few MS/MS success stories in identifying modified proteins (for example, Payne et al., 1991 identified phosphorylation sites of mitogen-activated protein kinase). The computational analysis of modified peptides was pioneered by Mann and Wilm, 1994 [13] and Yates et al., 1995 [17], [18]. The problem is particularly important since mass-spectrometry techniques sometimes introduces chemical modifications to native peptides and make these peptides "invisible" for database search programs. Mann and Wilm, 1994 [13] use a clever combination of partial *de novo* algorithm and database search in their *Peptide Sequence Tag* approach. Peptide sequence tag is a short run of clearly identifiable sequence ions that is used to reduce the search to the peptides containing this tag. This approach was successful in many applications, including identifying possible proteins in the apoptotic pathway (Shevchenko et al., 1997 [15]), but no information about its limitations and error rates for *mutation-tolerant* search is available. Yates et al.,1995 [17] suggested an exhaustive search approach that is to (implicitly) generate a virtual database of all modified peptides for a small set of modifications and to match the spectrum against this virtual database. Yates et al. 1995 [17] noted that it leads to a large combinatorial problem, even for a small set of modifications types and indicated that extending this approach to a larger set of modifications is an open problem. Another limitation is that extremely bulky modifications such as glycosylation disrupt the fragmentation pattern and would not be amenable to analysis by this method.

Mutation-tolerant database search in mass-spectrometry can be formulated as follows: given an experimental spectrum, find a peptide that matches the spectrum the best among the peptides that are at most k mutations apart from a database peptide. The problem is solved for $k = 0$ (any MS/MS database search program). MS-Tag software includes a program for $k = 1$ (Clauser et al , 1999 [3]) but the problem is unsolved for $k > 1$. It indicates that the current MS/MS database search programs are unable to detect peptides that are more than $5-10\%$ dissimilar, a rather narrow range.

We have developed a mutation-tolerant MS/MS database search and software, PEDANTA, to identify spectra of related peptides that differ by multiple mutations/modifications. PEDANTA reveals potential peptide modifications without exhaustive search and therefore does not require generating virtual database of modified peptides. We introduce a new measure of spectral similarity that is used to develop an efficient algorithm for mutation-tolerant database search and pairwise comparison of uncharacterized experimental spectra. The spectrum-to-spectrum comparison turned out to be a powerful method for obtaining spectra of interest from a large set of spectra. In particular, using our spectral similarity algorithm in conjunction with high-throughput tandem mass spectrometry, we have been able to determine possible phosphorylation sites for Chk1 kinase (Funari et al., 1997 [10]), a protein known to function in G2/M cell cycle regulation (this work will be described elsewhere).

2 Peptide identification problem

Let A be the set of amino acids with molecular masses $m(a)$, $a \in A$. A *peptide* $P = p_1 \dots p_n$ is a sequence of amino acids, the (parent) mass of peptide P is $m(P) = \sum m(p_i)$. A *partial peptide* $P' \subset P$ is a substring $p_i \dots p_j$ of P of mass $\sum_{i \le t \le j} m(p_t)$.

Peptide fragmentation in a *tandem mass spectrometer* can be characterized by a set of numbers $\Delta = \{\delta_1, \dots, \delta_k\}$ representing *ion-types*. A δ-*ion* of a partial peptide $P' \subset P$ is such modification of P' that has mass $m(P') - \delta$. For tandem mass spectrometry, *theoretical* spectrum of peptide P can be calculated by subtracting all possible ion-types $\delta_1, \dots, \delta_k$ from the masses of all partial peptides of P (i.e. every partial peptide generates k masses in the theoretical spectrum). An (experimental) spectrum $S = \{s_1, \dots, s_m\}$ is a set of masses of (fragment) ions. A *match* between spectrum S and peptide P is the number of masses that experimental and theoretical spectra have in common (shared peaks count). Dancik et al, 1999 [5] addressed the following

Peptide sequencing problem. Given spectrum S, the set of ion types Δ, and the mass m find a peptide of mass m with the maximal match to spectrum S.

Denote partial *N-terminal* peptide $p_1, \dots, p_i$ as P_i, and partial *C-terminal* peptide $p_{i+1}, \dots, p_n$ as P_i^-, $i = 1, \dots, n$. In practice MS/MS spectrum consists mainly of some of δ-ions of partial N-terminal and C-terminal peptides. For example, in the case of ion-trap mass spectrometer the most frequent N-terminal ions are b-*ions* (b_i corresponds to P_i with $\delta = -1$) and the most frequent C-terminal ions are y-ions (y_i corresponds to P_i^- with $\delta = 19$). Also, instead of the shared peaks count, the existing database search and *de novo* algorithms use more sophisticated objective functions (like weighted shared peaks count). We study the following

Peptide identification problem. Given a database of peptides, spectrum S, the set of ion types Δ, and parameter k find a peptide with the maximal match to spectrum S that is at most k mutations/modifications apart from a database entry.

The major difficulty in peptide identification problem comes from the fact that very similar peptides P_1 and P_2 may have very different spectra S_1 and S_2. Our goal is to define a notion of spectral similarity that correlates well with sequence similarity In other words, if P_1 and P_2 are a few substitutions/insertions/deletions/modifications apart, the spectral similarity between S_1 and S_2 should be high. Most existing database search programs are based on the shared peaks count that is, of course, an intuitive measure of spectral similarity. However, this measure diminishes very quickly as the number of mutations increases thus leading to limitations in detecting similarities in MS/MS database search. Moreover, there are many correlations between spectra of related peptides and only the small portion of them is captured by the "shared peaks" count. One can demonstrate that the "shared peaks" count captures roughly only $\frac{1}{2(k+1)}$ of the correlations between spectra of peptides that are k mutations apart. PEDANTA captures *all* correlations between related spectra for any k and handles the cases when mutations in the peptide significantly change the fragmentation pattern. For example, replacing amino acids like H, K, R, P may dramatically alter the fragmentation. Even in an extreme case like the one when a single mutation changes the fragmentation pattern from, let's say "only b-ions" to "only y-ions", PEDANTA still reveals the similarity between the corresponding spectra.

3 Spectral Convolution

Let S_1 and S_2 be two spectra. Define *spectral convolution* $S_2 \ominus S_1 = \{s_2 - s_1 : s_1 \in S_1, s_2 \in S_2\}$ and let $(S_2 \ominus S_1)(x)$ be the multiplicity of element x in this set. In other words, $(S_2 \ominus S_1)(x)$ is the number of pairs $s_1 \in S_1, s_2 \in S_2$ such that $s_2 - s_1 = x$. If $M(P)$ is the parent mass of peptide P with the spectrum S, then $S^R = M(P) - S$ is the *reversed spectrum* of S (every b-ion (y-ion) in S corresponds to y-ion (b-ion) in S^R). The *reversed spectral convolution* $(S_2 \ominus S_1^R)(x)$ is the number of pairs $s_1 \in S_1, s_2 \in S_2$ such that $s_2 + s_1 - M(P) = x$.

To illustrate the idea of the approach, consider two copies P_1 and P_2 of the same peptide. The number of peaks in common between S_1 and S_2 (shared peaks count) is the value of $S2 \ominus S1$ at $x = 0$. Most current MS/MS database search algorithms implicitly attempt to find a peptide P in the database that maximizes $S_2 \ominus S_1$ at $x = 0$, where S_2 is an experimental spectrum and S_1 is a theoretical spectrum of peptide P. However, if we start introducing k mutations in P_2 as compared to P_1, the value of $S2 \ominus S1$ at $x = 0$ quickly diminishes. As a result, the discriminating power of the "shared peaks" count falls down significantly at $k = 1$ and almost disappears at $k > 1$.

The new ingredient of our approach is an observation that peaks in spectral convolution allow one to detect mutations/modifications without exhaustive search Let P_2 differ from P_1 by the only mutation ($k = 1$) with amino acid difference $\delta = M(P_2) - M(P_1)$. In this case $S_2 \ominus S_1$ is expected to have two approximately equal peaks at $x = 0$ and $x = \delta$. If the mutation ocurrs at position t in the peptide then the peak at $x = 0$ corresponds to b_i-ions for $i < t$ and y_i-ions for $i \geq t$ The peak at $x = \delta$ corresponds to b_i-ions for $i \geq t$ and y_i-ions for $i < t$. A mutation in P_2 that changes $M(P_1)$ by δ also "mutates" the spectrum S_2 by shifting some peaks by δ. As a result, the number of shared peaks between S_1 and "mutated" S_2 may increase as compared to the number of shared peaks between S_1 and S_2. This increase is bounded by $(S_2 \ominus S_1)(\delta)$ and $(S_2 \ominus S_1)(0) + (S_2 \ominus S_1)(\delta)$ is an upper bound on the number of shared peaks between S_1 and "mutated" S_2.

The other set of correlations between spectra of mutated peptides is captured by the reverse spectral convolution $S_2 \ominus S_1^R$ reflecting the pairings of N-terminal and C-terminal ions (see Dancik et al., 1999 [5] for applications of reverse spectral convolution for parent mass computing) $S_2 \ominus S_1^R$ is expected to have two peaks at the *same* positions 0 and δ.

Now assume that P_2 and P_1 are two substitutions apart, one with mass difference δ_1 and another with $\delta - \delta_1$. These mutations generate two new peaks in the spectral convolution (at $x = \delta_1$ and at $x = \delta - \delta_1$). For uniform distribution of mutations in a random peptide, the ratio of the expected heights of the peaks at $0, \delta, \delta_1, \delta - \delta_1$ is $2 : 2 . 1 : 1$.

Different fragment ions contribute to different peaks but short fragment ions contribute mainly to peaks at 0 and δ. Since short fragment ions are frequently missing from the spectra (for ion-trap mass-spectrometers) the heights of peaks at 0 and δ are more in line with the heights of the peaks at δ_1 and $\delta - \delta_1$ in practice. Therefore, the "shared peaks" count ignores 75% of correlations in the related spectra for $k = 1$ and even more for $k > 1$.

To increase the signal/noise ratio we combine the peaks in spectral and reverse spectral convolution

$$S = S_2 \ominus S_1 \; + \; S_2 \ominus S_1^R$$

Further we combine the peaks at 0 and δ (as well as at δ_1 and $\delta - \delta_1$) by introducing the *shift function*

$$F(x) = \frac{1}{2}(S(x) + S(\delta - x))$$

Note that $F(x)$ is symmetric around the axis $x = \frac{\delta}{2}$ with $F(0) = F(\delta)$ and $F(\delta_1) = F(\delta - \delta_1)$. We are interested in the peaks of $F(x)$ for $x \geq \frac{\delta}{2}$.

Define $x_1 = \delta = M(P_2) - M(P_1)$ and $y_1 = F(\delta) = F(0)$. Let $y_2 = F(x_2), y_3 = F(x_3), \ldots, y_k = F(x_k)$ be $k-1$ largest peaks of $F(x)$ for $x \geq \delta/2$ and $x \neq \delta$. Define

$$SIM_k(S_1, S_2) = \sum_{i=1}^{k} y_i$$

as an estimate for the similarity between spectra S_1 and S_2 under the assumption that the corresponding peptides are k mutations apart. SIM_k is usually the overall height of k highest peaks of the shift function. For example, $SIM_1(S_1, S_2) = y_1$ is an upper bound for the number of shared peaks between S_1 and "mutated" S_2 if $k = 1$ mutation in P_2 is allowed. Note the difference in use of the spectral convolution in our approach (analysis of top peaks to reveal mutations) and the cross-correlation function from Eng et al., 1994 [7].

In a more practical version of the same definition, $x_2, x_3, \ldots, x_k$ are restricted to valid mass shifts, i.e. to the values that correspond to amino acid mass differences (substitutions), amino acid masses (deletions/insertions) or to the mass differences corresponding to potential amino acid modifications

The definition of SIM_k given above treats all fragment ions equally without attempting to take into account intensities and to score the major ions (like b) and minor ions (like $b - H_2O$) according to their propensities. To account for intensities and propensities one can assign a score to every peak in the experimental spectrum that gives more weight to the high-intensity peaks and peaks explained by multiple fragment-ions See Dancik et al., 1999 [5] for offset frequency function approach to such scoring.

The masses of amino acids present in peptides may generate "false" peaks in spectral convolution. Other false peaks may correspond to $\delta = 18$ and $\delta = 17$ (loss of H_2O or NH_3), or more precisely, every peak in spectral convolution may have a twin peak shifted by the mass of H_2O or NH_3. Below we describe even more serious limitation of the shift function

Let

$$S = \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$$

be a spectrum of peptide P and assume for simplicity that P produces only b-ions. Let

$$S' = \{10, 20, 30, 40, 50, 55, 65, 75, 85, 95\}$$

and

$$S'' = \{10, 15, 30, 35, 50, 55, 70, 75, 90, 95\}$$

be two theoretical spectra corresponding to peptides P' and P'' from the database. Which peptide (P' or P'') fits spectrum S the best? The "shared peaks" count does not allow one to answer this question since both S' and S'' have 5 peaks in common with S. Moreover, the spectral convolution also does not answer this question since both $S \ominus S'$ and $S \ominus S''$ (and corresponding shift functions) reveal the strong peaks of the same height at 0 and 5. It suggests

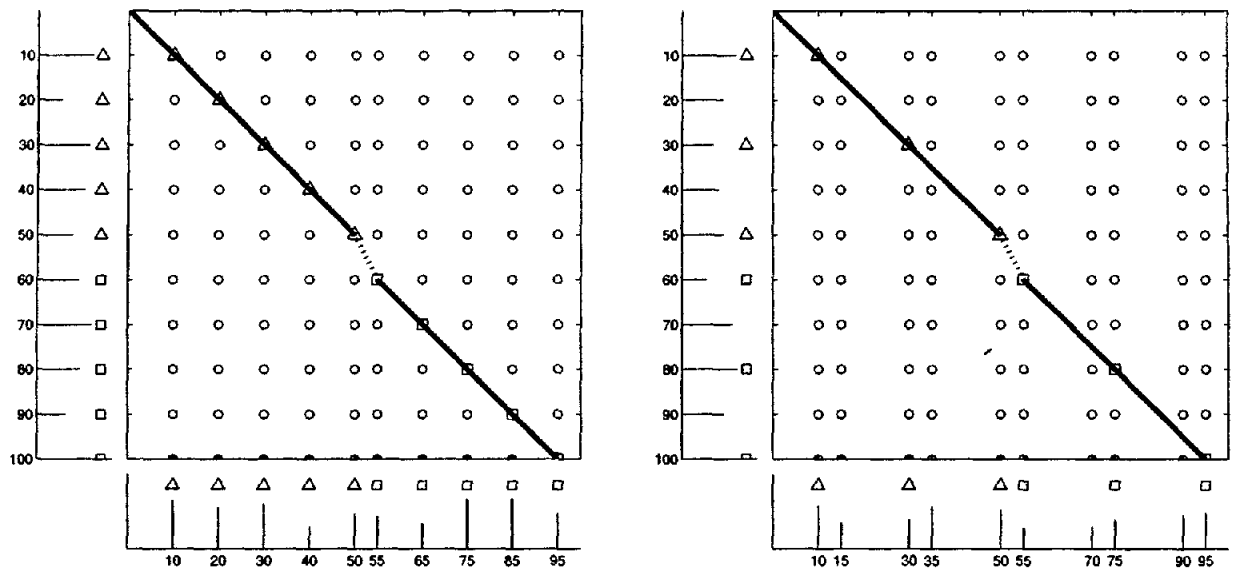

Figure 1: Spectrum S can be transformed into S' by a single mutation and $D(1) = 10$ (left matrix). Spectrum S cannot be transformed into S'' by a single mutation and $D(1) = 6$ (right matrix).

that both P' and P'' can be obtained from P by a single mutation with mass difference 5. However, a more careful analysis shows that although this mutation can be realized for P' by introducing a shift 5 after mass 50, it cannot be realized for P''. The major difference between S' and S'' is that the matching positions in S' come in clumps while the matching positions in S'' don't This important property of related spectra was not captured by spectral convolution and was overlooked in the previous studies of MS/MS database search. Below we describe the spectral alignment approach to address this problem.

4 Spectral Alignment

Let $A = \{a_1, \ldots a_n\}$ be an ordered set of natural numbers $a_1 < a_2 \ldots < a_n$. *A shift* Δ_i transforms A into $\{a_1, \ldots a_{i-1}, a_i + \Delta_i, \ldots, a_n + \Delta_i\}$. We consider only the shifts that do not change the order of elements, i.e the shifts with $\Delta_i \geq a_{i-1} - a_i$. Given sets $A = \{a_1, \ldots a_n\}$ and $B = \{b_1, \ldots b_m\}$ we want to find a series of k shifts of A that make A and B as similar as possible. The k-similarity $D(k)$ between sets A and B is defined as the maximum number of elements in common between these sets after k shifts. For example, a shift -5_6 transforms $S = \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$ into $S' = \{10, 20, 30, 40, 50, 55, 65, 75, 85, 95\}$ and therefore $D(1) = 10$ for these sets. The set $S'' = \{10, 15, 30, 35, 50, 55, 70, 75, 90, 95\}$ has 5 elements in common with S (the same as S') but there is no shift transforming S into S'' and $D(1) = 6$. Below we describe a dynamic programming algorithm for computing $D(k)$.

Define a *spectral product* $A \otimes B$ as $a_n \times b_m$ two-dimensional matrix with nm 1s corresponding to all pairs of indices (a_i, b_j) and remaining elements being zeroes. The number of 1s at the main diagonal of this matrix describes the "shared peaks count" between spectra A and B, or in another words, 0-similarity between A and B. Figure 1 shows the spectral product $S \otimes S'$ and $S \otimes S''$ for an example from the previous section. In both cases the number of 1s on the main diagonal is the same and $D(0) = 5$. The "δ-shifted peaks count" is the number of 1s on the diagonal $(i, i + \delta)$. The limitation of the shift function is that it considers diagonals separately without combining them into feasible mutation scenarios.

Define a directed graph with vertices corresponding to 1s in the spectral product and edges corresponding to pairs of vertices (i, j) and (i', j') with $i \leq i'$ and $j \leq j'$. *k-similarity* between spectra is defined as the maximum number of 1s on a path through this graph that uses at most $k + 1$ diagonals and *k-optimal spectral alignment* is defined as a path using these $k + 1$ diagonals For example, 1-similarity is defined by the maximum number of 1s on a path through the spectral product that uses at most two diagonals. Figure 1 reveals that the notion of 1-similarity allows one to find out that S is closer to S' than to S'' since in the first case the 2-diagonal path cover 10 ones (left matrix) versus 6 in the second case (right matrix). Figure 2 illustrates that the spectral alignment allows one to detect more and more subtle similarities between spectra by increasing k. Below we describe a dynamic programming algorithm for spectral alignment.

Let A_i and B_j be i-prefix of A and j-prefix of B correspondingly. Define $D_{i,j}(k)$ as the k-similarity between A_i and B_j such that the last elements of A_i and B_j are matched In other words, $D_{i,j}(k)$ is the maximum number of 1s on a path to (a_i, b_j) that uses at most $k + 1$ diagonals. We say that (i', j') and (i, j) are *co-diagonal* if $a_i - a_{i'} = b_j - b_{j'}$ and that $(i', j') < (i, j)$ if $i' < i$ and $j' < j$. To take care of the initial conditions we introduce a fictitious element $(0, 0)$ with $D_{0,0}(k) = 0$ and assume that $(0, 0)$ is co-diagonal with any other (i, j). The dynamic programming recurrency for

$D_{ij}(k)$ is

$$D_{ij}(k) = \max_{(i',j')<(i,j)} \begin{cases} D_{i'j'}(k)+1, & \text{if } (i',j') \text{ and } (i,j) \\ & \text{are co-diagonal} \\ D_{i'j'}(k-1)+1, & \text{otherwise} \end{cases}$$

The k-similarity between A and B is given by $D(k) = \max_{ij} D_{ij}(k)$.

The described dynamic programming algorithm for spectral alignment is rather slow (running time $O(n^4k)$ for n-element spectra) and below we describe $O(n^2k)$ algorithm for solving this problem. Define $diag(i,j)$ as the maximal co-diagonal pair of (i,j) such that $diag(i,j) < (i,j)$. In other words, $diag(i,j)$ is the position of previous 1 on the same diagonal as (a_i, b_j) or $(0,0)$ if such position does not exists. Define

$$M_{ij}(k) = max_{(i',j')\le(i,j)} D_{i'j'}(k)$$

Then the recurrency for $D_{ij}(k)$ can be re-written as

$$D_{ij}(k) = \max \begin{cases} D_{diag(i,j)}(k)+1, \\ M_{i-1,j-1}(k-1)+1 \end{cases}$$

The recurrency for $M_{ij}(k)$ is given by

$$M_{ij}(k) = \max \begin{cases} D_{ij}(k) \\ M_{i-1,j}(k) \\ M_{i,j-1}(k) \end{cases}$$

The described transformation of dynamic programming graph is achieved by introducing horizontal and vertical edges that provide switching between diagonals (Figure 3). The score of the path is the number of 1s on this path while k corresponds to the number of switches (number of used diagonals minus 1).

5 Aligning Peptide against Spectrum

The simple description above hides many details that make the spectral alignment problem difficult. These details include simultaneous analysis of N-terminal and C-terminal ions, taking into account the intensities and charges, analysis of minor ions, etc.

Spectra are usually a combination of an increasing (N-terminal ions) and a decreasing (C-terminal ions) number series. These series form two diagonals in the spectral product $S \otimes S$, the main diagonal and the perpendicular diagonal that corresponds to pairings of N-terminal and C-terminal ions. The described algorithm does not capture this specifics and deals with the main diagonal only.

To combine N-terminal and C-terminal series together we work with $(S_1 \cup S_1^R) \otimes (S_2 \cup S_2^R)$ where S^R is the reversed spectrum of peptide P. This transformation creates a "b-version" for every y-ion and "y-version" for every b-ion thus increasing noise (since every noisy peak is propagated twice) Another and even more serious difficulty is that every 1 in the spectral product will have a reversed twin and only one of these twins should be counted in the feasible spectral alignment. Dancik et al., 1999 [5] demonstrated that ignoring this problem may lead to infeasible solutions and formulated *anti-symmetric path problem* that addresses this issue In the later work, Dancik and Pevzner, 1999 [6] suggested a polynomial algorithm for anti-symmetric path problem.

The described algorithm also does not capture all the relevant details in the case of the "sequence against the

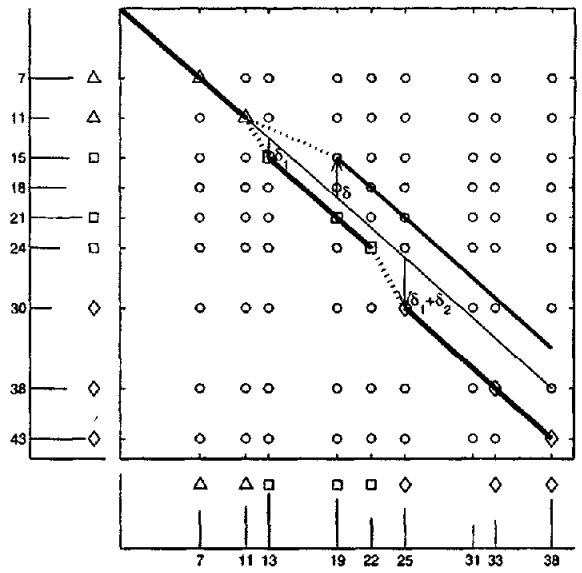

Figure 2. Aligning Spectra. The "shared peaks" count reveals only $D(0) = 3$ matching peaks on the main diagonal while spectral alignment reveals more hidden similarities between spectra ($D(1) = 5$ and $D(2) = 8$) and detects the corresponding mutations.

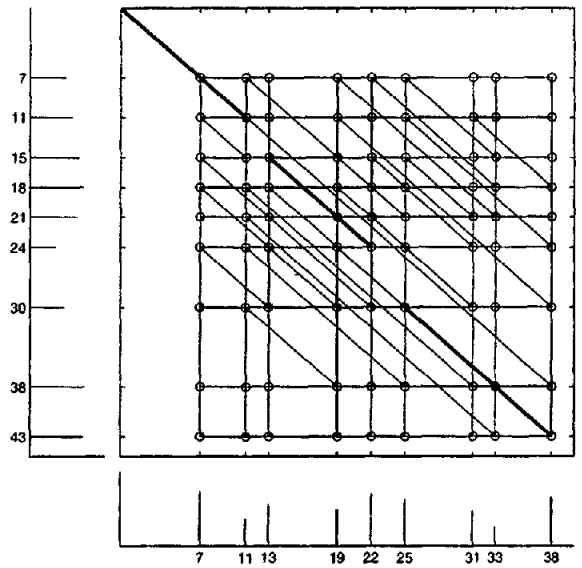

Figure 3. Modification of dynamic programming graph leads to a fast spectral alignment algorithm.

spectrum" comparison. In this case the horizontal and vertical arcs in the dynamic programming graph (Fig. 3) are limited by the possible shifts reflecting mass differences between amino acid participating in the mutation. Let $P = p_1 \ldots p_n$ be a peptide that we compare with the spectrum $S = \{s_1, \ldots, s_m\}$. d-prefix of spectrum S contains all peaks of S with $s_i \leq d$. We introduce new variable $H_{i,d}(k)$ that describes the "best" transformation of the i-prefix of peptide P into d-prefix of spectrum S with at most k substitutions in P_i. More precisely, $H_{i,d}(k)$ describes the number of 1s on the optimal path with k shifts between diagonals from $(0,0)$ to the position (i,d) of the properly defined "peptide versus spectrum" $P \otimes S$ matrix. Also, for the sake of simplicity, assume that the theoretical spectrum of P contains only b-ions.

Let $H_{i,d}(k)$ be the "best" transformation of P_i into S_d with k substitutions (i e. a transformation that uses maximum number of 1s on a path with at most k shifts between diagonals). However, in this case the jumps between diagonals are not arbitrary but are restricted by mass differences of mutated amino acids (or mass differences corresponding to chemical modifications). Below we describe the dynamic programming algorithm for the case of substitutions (deletions/insertions and modifications lead to similar recurrencies). Define $x(d) = 1$ if $d \in S$ and $x(d) = 0$ otherwise. Then $H_{i,d}(k)$ is described by the following recurrency ($m(a)$ is the mass of amino acid a):

$$H_{i,d}(k) = \max \begin{cases} H_{i-1,d-m(p_i)}(k) + x(d) \\ \max_{a=1,20} H_{i,d-(m(a)-m(p_i))}(k-1) \end{cases}$$

The computational complexity of the above algorithms is $O(ns_mk)$ assuming that the spectrum and masses of amino acids are integers. It is important to notice that the computations in the above algorithm should go in the increasing order of k.

6 Conclusion

We described a mutation-tolerant database search approach that is based on a new notion of spectral similarity. An alternative to this method is *de novo* interpretation followed by a BLAST-like database similarity search as proposed by Taylor and Johnson, 1997 [16] and Clauser (personal communication) This approach shows a hope for mutation-tolerant searches but is unlikely to succeed for modification-tolerant searches since *de novo* reconstruction of modified peptides remains an open problem.

PEDANTA has been tested on both experimental and simulated data These tests demonstrated that PEDANTA is very efficient for mutation-tolerant database search with up to two mutations even for relatively poor spectra. PEDANTA also captures many related spectra for $k = 3$ but in this case a two-stage procedure with a more accurate objective function is required. The results of the tests will be described elsewhere (Mulyukov and Pevzner, 1999 [14])

7 Acknowledgements

We are indebted to Karl Clauser for many critical comments that greatly improved this paper. We also grateful to Terry Addona and Jim Vath for the discussion on applications of the method.

References

[1] C. Bartels Fast algorithm for peptide sequencing by mass spectroscopy *Biomed. Environ. Mass Spectrom* , 19:363–368, 1990

[2] K. Biemann and H. A. Scoble Characterization of tandem mass spectrometry of structural modifications in proteins *Science*, 237.992–998, 1987

[3] K R. Clauser, P. R Baker, and A. L Burlingame The role of accurate mass measurement (+/- 10ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem*, 71.2871–2882, 1999

[4] K. R. Clauser, P.R. Baker, and A.L. Burlingame. In *44th ASMS Conference on Mass Spectrometry and Allied Topics, Portland, Oregon, May 12-16*, page 365, Portland, OR, May 12-16 1996

[5] V. Dancik, T. Addona, K. Clauser, J Vath, and P.A Pevzner. De novo peptide sequencing via tandem mass spectrometry *J. Comp. Biology*, 6:327–342, 1999

[6] V. Dancik and P A. Pevzner. Anti-symmetric path problem in tandem mass-spectrometry *(in preparation)*, 1999

[7] J Eng, A. McCormack, and J Yates An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc. Mass Spectom* , 5:976–989, 1994.

[8] D. Fenyo, J. Qin, and B.T Chait Protein identification using mass spectrometric information *Electrophoresis*, 19 998–1005, 1998

[9] J. Fernández-de Cossío, J Gonzales, and V Besada. A computer program to aid the sequencing of peptides in collision-activated decomposition experiments *CABIOS*, 11:427–434, 1995.

[10] B Funari, N. Rhind, and P. Russell. Cdc25 mitotic inducer targeted by Chk1 DNA damage checkpoint kinase *Science*, 277.1495–1497, 1997.

[11] A Gooley and N. Packer The importance of co- and post-translational modifications in proteome projects In W. Wilkins, K Williams, R Appel, and D Hochstrasser, editors, *Proteome Research: New Frontiers in Functional Genomics*, pages 65–91. Springer, 1997

[12] R. J Johnson and K Biemann Computer program (seqpep) to aid in the interpretation of high-energy collision tandem mass spectra of peptides *Biomed. Env. Mass Spectrom* , 18 945–957, 1989.

[13] M. Mann and M. Wilm. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.*, 66.4390–4399, 1994

[14] Z. Mulyukov and P.A Pevzner Efficiency of mutation-tolerant database search with tandem mass spectra *(in preparation)*, 1999

[15] A. Shevchenko, M Wilm, and M. Mann. Peptide mass spectrometry for homology searches and cloning of genes. *J Protein Chemistry*, 5 481–490, 1997

[16] J A Taylor and R S Johnson. Sequence database searches via *de novo* peptide sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 11.1067–1075, 1997.

[17] J Yates, J. Eng, A McCormack, and D. Schieltz. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database *Anal. Chem.*, 67 1426–1436, 1995

[18] J. R Yates, J. K Eng, and A L McCormack. Mining genomes Correlating tandem mass-spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal Chem.*, 67 3202–3210, 1995.