A New Approach to Protein Identification

Nuno Bandeira, Dekel Tsur, Ari Frank, and Pavel Pevzner

University of California, San Diego, Dept. of Computer Science and Engineering, 9500 Gilman Drive, La Jolla, CA 92093, USA. {nbandeir,dtsur,amfrank,ppevzner}@cs.ucsd.edu.

Abstract. Advances in tandem mass-spectrometry (MS/MS) steadily increase the rate of generation of MS/MS spectra and make it more computationally challenging to analyze such huge datasets. As a result, the existing approaches that compare spectra against databases are already facing a bottleneck, particularly when interpreting spectra of post-translationally modified peptides. In this paper we introduce a new idea that allows one to perform MS/MS database search ... without ever comparing a spectrum against a database. The idea has two components: experimental and computational. Our experimental idea is counter-intuitive: we propose to intentionally introduce chemical damage to the sample. Although it does not appear to make any sense from the experimental perspective, it creates a large number of "spectral pairs" that, as we show below, open up computational avenues that were never explored before. Having a spectrum of a modified peptide paired with a spectrum of an unmodified peptide, allows one to separate the prefix and suffix ladders, to greatly reduce the number of noise peaks, and to generate a small number of peptide reconstructions that are very likely to contain the correct one. The MS/MS database search is thus reduced to extremely fast pattern matching (rather than time-consuming matching of spectra against databases). In addition to speed, our approach provides a new paradigm for identifying post-translational modifications.

1 Introduction

Most protein identifications today are performed by matching spectra against databases using programs like SEQUEST [7] or MASCOT [15]. While these tools are invaluable, they are already too slow for matching large MS/MS datasets against large protein databases. Moreover, the recent progress in mass-spectrometry instrumentation (a single LTQ-FT mass-spectrometer can generate 100,000 spectra per day) may soon make them obsolete. Since SEQUEST compares every spectrum against every database peptide, it will take a cluster of about 60 processors to analyze the spectra produced by a single such instrument in real time (if searching through the Swiss-Prot database). If one attempts to perform a time-consuming search for post-translational modificatoions, the running time may further increase by orders of magnitude. We argue that new solutions are needed to deal with the stream of data produced by shotgun proteomics projects. Beavis et al, 2004 [5] and Tanner et al., 2005 [20] recently developed X!Tandem and InsPecT algorithms that prune (X!Tandem) and filter (InsPecT) databases to speed-up the search. However, these tools still have to compare every spectrum against a (smaller) database.

In this paper we explore a new idea that allows one to perform MS/MS database search without ever comparing a spectrum against a database. The idea has two components: experimental and computational. Our experimental idea, while counter-intuitive, is trivial to implement. We propose to slightly change the experimental protocol by intentionally introducing chemical damage to the sample and generating many modified peptides. The current protocols try to achieve the opposite goal of minimizing the chemical damage since (i) modified peptides are difficult to interpret and (ii) chemical adducts do not provide any useful information. Nevertheless, the existing experimental protocols unintentionally generate many chemical mod-

ifications (sodium, potassium, Fe(III), etc.)[1] and below we show that existing MS/MS datasets often contain modified versions for many peptides. In addition, even in a sample without chemical damage, exopeptidases routinely create a variety of peptides that differ from each other by a deletion of terminal amino acids.

From the experimental perspective, subjecting a sample to chemical damage does not make any sense.[2] However, from the computational perspective, it creates a large number of "spectral pairs" that, as we show below, open up computational avenues that were never explored before. Having a pair of spectra (one of a modified and another of an unmodified peptide) allows one to separate the prefix and suffix mass ladders, to greatly reduce the number of noise peaks, and to generate a small number of peptide reconstructions that are very likely to contain the correct one. In difference from our recent approach to generating *covering sets* of short 3–4 amino acid *tags* (Frank et al., 2005 [9], Tanner et al., 2005 [20]), this approach generates a small covering set of *peptides* 7–9 amino acids long. This set typically has a single perfect hit in the database that can be instantly found by hashing and thus eliminates the need to ever compare the spectrum against the database.[3]

Let $S(P)$ and $S(P^*)$ be spectra of an unmodified peptide P and its modified version P^* (spectral pair). The crux of our computational idea is a simple observation that a "database" consisting of a single peptide P is everything one needs to interpret the spectrum $S(P^*)$. If one knows P then there is no need to scan $S(P^*)$ against the database of all proteins! Of course, in reality one does not know P since only $S(P)$ is available. Below we show that the spectrum $S(P)$ is nearly as good as the peptide P for interpreting $S(P^*)$ thus eliminating the need for database search. This observation opens the possibility of substituting MS/MS database search with finding spectral pairs and further interpreting the peptides that produced them. Below we show that these problems can be solved using a variation of the spectral alignment approach [16]. We further show how to transform the spectral pair (S_1, S_2) into virtual spectra $S_{1,2}$ and $S_{2,1}$ of extremely high quality; with nearly perfect b and y ion separation and the number of noisy peaks reduced twelvefold, these spectra (albeit virtual) are arguably the highest quality spectra mass-spectrometrists ever saw.

2 Dataset

We describe our algorithm and illustrate the results using a sample of MS/MS spectra from IKKb protein. The IKKb dataset consists of 45,500 spectra acquired from a digestion of the inhibitor of nuclear factor kappa B kinase beta subunit (IKKb protein) by multiple proteases, thereby producing overlapping peptides (spectra were acquired on a Thermo Finigan LTQ mass spectrometer). The activation of the inhibitor kappaB kinase (IKK) complex and its relationships to insulin resistance were the subject of recent intensive studies. The IKK complex represents an ideal test case for algorithms that search for post-translationally modified (PTM) peptides. Until recently, phosphorylations were the only known PTMs in IKK, which does not explain mechanisms of signaling and activation/inactivation of IKK by over 200 different stimuli, including cytokines, chemicals, ionization and UV radiation, oxidative stress, etc. It is likely that different

[1] Hunyadi-Goulyas and Medzihradszky, 2004 [10] give a table of over 30 common chemical adducts that are currently viewed as annoyances.

[2] Probably the easiest way to chemically damage the sample is to warm it up in urea solution or to simply bring it into mildly acidic pH and add a hefty concentration of hydrogen peroxide. See Levine et al., 1996 [14] for an example of a slightly more involved protocol that generates samples with desired extent of oxidation in a controlled fashion. Also, to create a mixture of modified and unmodified peptides, one can split the sample in half, chemically damage one half, and combine both halves together again.

[3] We remark that the Peptide Sequence Tag approach reduces the number of considered peptides but does not eliminates the need to match spectra against the *filtered* database. For example, Tanner et al., 2005 [20] describe a dynamic programming approach for matching spectra against a filtered database.

stimuli use different mechanisms of signaling involving different PTM sites. Revealing the combinatorial code responsible for PTM-controlled signalling in IKK remains an open problem.

The IKKb dataset was extensively studied in Tanner et al., 2005 [20] and Tsur et al., 2005 [21] resulting in 11760 identified spectra and 1154 annotated peptides (p-value $\leq$0.05). This IKKb sample presents an excellent test case for our protocol since 77% of all peptides in this sample have spectral pairs even without intentionally subjecting the sample to chemical damage. 639 out of 1154 annotated peptides are modified. 448 out of 639 modified peptides have an unmodified variant. 208 out of 515 unmodified peptides have a modified version, and 413 out of 515 unmodified peptides have either a modified version or a prefix/suffix peptide in the sample. The sample contains 571 peptides with 3 or more spectra (345 unmodified and 226 modified), 191 peptides with 2 spectra (71 unmodified and 120 modified) and 392 peptides with a single spectrum (99 unmodified and 293 modified). The dataset has not been manually validated and the unusually high proportion of modified peptides with a single spectrum as compared to peptides annotated by multiple spectra may be an indication that some annotations of peptides explained by a single spectrum may be incorrect.

3 Detecting spectral pairs

3.1 Clustering spectra

Clustering multiple spectra of the same peptide achieves a twofold goal: (i) the consensus spectrum of a cluster contains much fewer noise peaks than the individual spectra, and (ii) clustering speeds up and simplifies the search for spectral pairs. The clustering step capitalizes on the fact that true peaks consistently occur in multiple spectra from the same peptide, while noise peaks do not. Our clustering approach follows Bandeira et al., 2004 [1] with some improvements outlined below. We first transform every spectrum into its scored version that substitutes peak intensities with log likelihood scores. Any scoring used in *de novo* peptide sequencing algorithms can be used for such transformation (we have chosen to use scoring from Frank and Pevzner [8]). We also transform every spectrum into a PRM spectrum (see [1]).

Bandeira et al. [1] use a spectral similarity measure to decide whether two spectra come from the same peptide. While spectral similarity largely succeeds in identifying related spectra, it may in some cases pair non-related spectra. Although such false pairings are rare, they may cause problems if they connect two unrelated clusters. To remove false pairs we use a heuristic approach from Ben-Dor et al. [2]. This clustering procedure resulted in 567 clusters representing 98% of all unmodified and 96% of all modified peptides with three or more spectra in the original sample.

Each cluster of spectra is then collapsed into a single *consensus spectrum* that contains peaks present in at least k spectra in the cluster. The parameter k is chosen in such a way that the probability of seeing a peak in k spectra by chance is below 0.01.[4] We further sum up the scores of matching peaks to score the peaks in the consensus spectrum. As shown in Table 1, the resulting consensus spectra have unusually high signal-to-noise ratio (the number of unexplained peaks in the consensus spectra is reduced by a factor of 2.5). We also observed some consistently co-occurring unexplained peaks possibly due to co-eluting peptides or unexplained fragment ions (e.g., internal ions). After clustring we end up with 567 consensus spectra (that cover 93% of all individual spectra) and 862 unclustered spectra.

[4] We model the noise peak generation as a Bernoulli trial and the occurrence of k matching peaks in a cluster of n spectra as random variable with a Binomial distribution.

Type of spectra		*#Explained*			*#Unexplained*	*#Total*	*signal-to-noise ratio*
		b	y	*Satellite*			
Single spectra	# peaks:	9.48	9.26	20.07	35.25	74.05	**0.27**
(11760 spectra)	% peaks:	13%	13%	26%	48%		
	% score:	28%	28%	19%	25%		
Consensus spectra	# peaks:	9.47	9.39	10.42	13.74	43.06	**0.69**
(567 spectra)	% peaks:	22%	22%	24%	32%		
	% score:	37%	36%	13%	14%		
Spectral pairs $S^b_{i,j}$	# peaks:	6.47	0.2	0.38	1.69	8.64	**3.83**
(1569 pairs)	% peaks:	75%	2%	4%	19%		
	% score:	87%	2%	4%	7%		
Star spectra	# peaks:	8.38	0.52	0.92	2.9	12.72	**2.89**
(745 stars)	% peaks:	66%	4%	7%	23%		
	% score:	88%	3%	2%	7%		

Table 1. Statistics of single spectra, consensus spectra, spectral pairs, and star spectra. Satellite peaks include fragment ions correlated with b and y peaks ($b - H_2O$, $b - NH_3$, a, b^2, etc.). Signal-to-noise ratio is defined as $\frac{\#b-ions}{\#unexplained\ peaks}$. Spectral pairs separate prefix and suffix ladders and make interpretations of resulting spectra $S^b_{i,j}$ straightforward. Spectral stars further increase the number of b and y peaks in the resulting spectra. Note that b peaks are responsible for about 90% of the score in both paired and star spectra. The results are given only for the $S^b_{i,j}$ spectra since the $S^y_{i,j}$ spectra have the same statistics.

3.2 Spectral pairs

Peptides P_1 and P_2 form a *peptide pair* if either (i) P_1 differs from P_2 by a single modification/mutation, or (ii) P_1 is either a prefix or suffix of P_2[5]. Two spectra form a *spectral pair* if their corresponding peptides are paired. Although the peptides that give rise to a spectral pair are not known in advance, we show below that spectral pairs can be detected with high confidence using uninterpreted spectra.

For two spectra S_1 and S_2, the *spectral product* [16] of S_1 and S_2 is the set of points $(x, y) \in \mathbb{R}^2$ for every $x \in S_1$ and $y \in S_2$ (S_1 and S_2 are represented as sets of masses). Figure 1a shows the spectral product for the theoretical spectra of two peptides. The similarity between the two spectra is revealed by two diagonals (blue and red) in the spectral product.

[5] Condition (ii) can be viewed as a variation of (i) if one considers a pair of peptides differing by a few prefix/suffix residues as a single mutation (such variations are common in MS/MS samples). More generally, peptides P_1 and P_2 form a peptide pair if either (i) P_1 is a modified/mutated version of P_2, or (ii) P_1 and P_2 overlap. While our techniques also work for this generalization, we decided to limit our analysis to simple peptide pairs described above. We found that such simple pairs alone allow one to interpret most spectra. Adding pairs of spectra with more subtle similarities further increases the number of spectral pairs but slows down the algorithm.

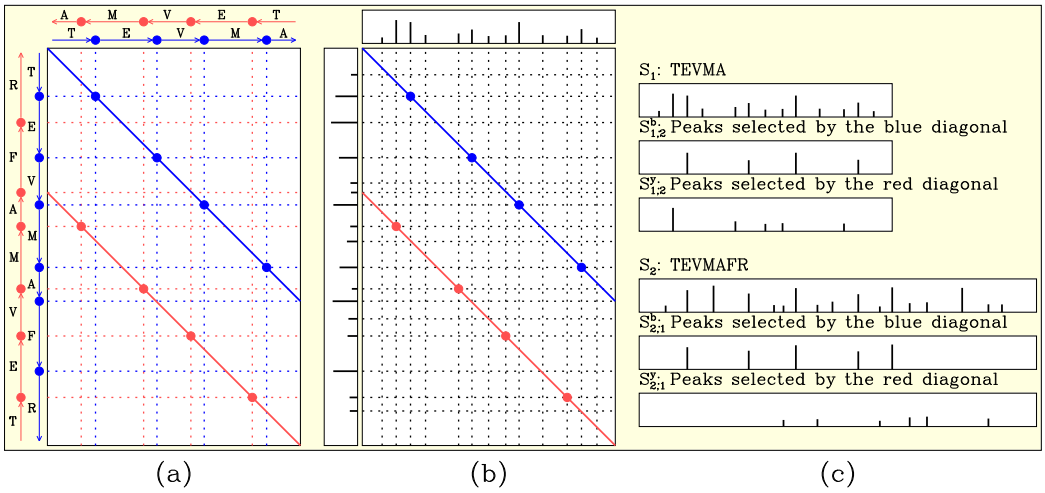

Fig. 1. Spectral product for the spectra of the peptides TEVMA and TEVMAFR. Figure (a) shows the spectral product for the theoretical spectra of these peptides (all points at the intersections between the vertical and horizontal lines). The blue (resp., red) circles correspond to matching b ions (resp., y ions) in the two spectra. The blue and red circles are located on the blue and red diagonals. Figure (b) shows the spectral product for uninterpreted spectra of the peptides TEVMA and TEVMAFR. The two diagonals in the spectral product matrix still reveal the points where peaks from the spectrum at the top match peaks from the spectrum on the left. Figure (c) illustrates how the blue and red diagonals define the spectra $S^b_{1,2}$ and $S^y_{1,2}$.

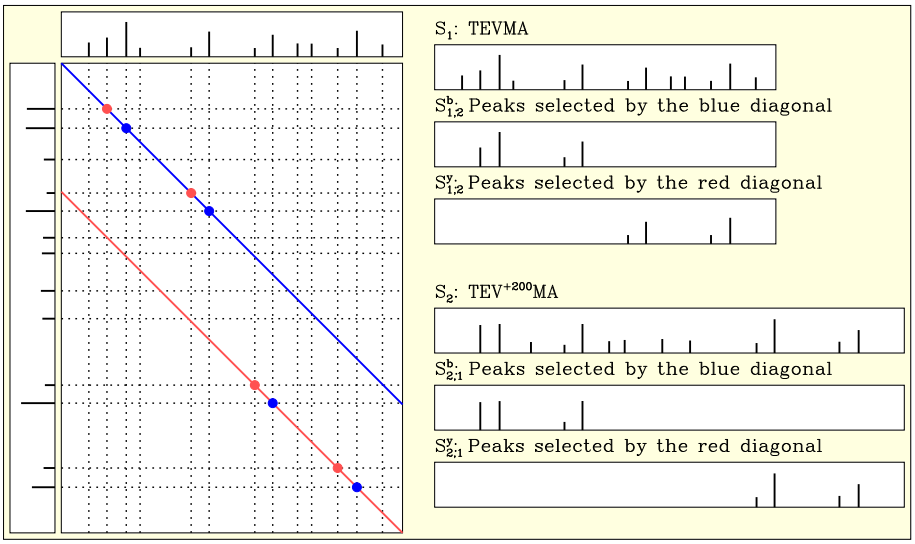

Fig. 2. Spectral product matrix for uninterpreted spectra with internal modification: The first spectrum corresponds to an unmodified peptide, and the second spectrum corresponds to a modified peptide. In these cases it is not appropriate to construct $S^b_{i,j}/S^y_{i,j}$ by simply selecting peaks on the diagonals - section 3.4 describes a more adequate algorithm for this purpose.

Figures 1b and 2 show pairs of uninterpreted spectra, denoted S_1 and S_2, and their spectral product. Although the "colors" of peaks are not known in this case, we take the liberty to name one diagonal blue and the other red. One can use circles (matching peak masses) on the blue diagonal to transform the original spectrum S_1 into a new spectrum $S^b_{1,2}$ (Figure 1c) with a much smaller number of peaks (a peak in S_1 is retained in $S^b_{1,2}$ only if it generates a circle on the blue diagonal). Similarly, one can transform S_1 into a spectrum $S^y_{1,2}$ using circles on the red diagonal. The peak scores in both spectra $S^b_{1,2}$ and $S^y_{1,2}$ are inherited from spectrum S_1. Similarly, the spectrum S_2 is transformed into spectra $S^b_{2,1}$ and $S^y_{2,1}$.[6]

Intuitively, if two spectra are unrelated, blue and red diagonals represent random matches and the number of circles appearing on these diagonals is small. Paired spectra, on the contrary, are expected to have many circles on these diagonals. Although this simple criterion (number of circles on two diagonals) would already allow one to roughly distinguish paired spectra from unrelated spectra, we describe below a more accurate test for finding spectral pairs.

3.3 Spectral pairs graph

The *correlation score* of spectra S_1 and S_2 is defined as the total score of all peaks in spectra $S^b_{1,2}$ and $S^y_{1,2}$: $score(S_1, S_2) = score(S^b_{1,2}) + score(S^y_{1,2})$. Similarly, $score(S_2, S_1) = score(S^b_{2,1}) + score(S^y_{2,1})$. We accept S_1 and S_2 as a putative spectral pair if both the ratio $\frac{score(S_1,S_2)}{score(S_1)}$ and $\frac{score(S_2,S_1)}{score(S_1)}$ exceed a predefined threshold (0.4 in examples below).

In addition to the correlation score test described above, we also use a test that takes into account the size of the MS/MS sample. The larger the set of spectra under consideration the larger the chance that a certain correlation score can be achieved by chance. To account for this phenomenon we assume that the correlation scores between unrelated spectra approximately follow a Gaussian distribution. Thus, a correlation score is only considered significant if the probability of this score appearing by chance is below 0.01.

The spectral pairs that satisfy both tests form the *spectral pairs graph* on the set of all spectra (Figure 3). The spectral pairs graph for the IKKb dataset has 43 connected components with 1021 vertices and 1569 edges. The small number of connected components is not surprising since overlapping peptides in this dataset can be assembled into a small number of contigs (an effect previously explored in the context of shotgun protein sequencing [1]). The combined filtering efficiency of these criteria allowed us to retain 78.4% of all correct spectral pairs at a precision level of 95% and find several different variants of most unmodified peptides. Table 1 describes the statistics of spectra $S_{i,j}$ and shows the dramatic increase in signal-to-noise ratio as compared to consensus spectra (let alone individual spectra). Moreover, the spectral pairs provide nearly perfect separation between prefix and suffix ladders thus making follow up interpretation straightforward. When compared to EigenMS's [3] average performance on single LTQ MS/MS spectra, spectral pairs reduce the contamination of suffix peaks in prefix ladders (and vice-versa) from their reported level of 11% to only 2%.

3.4 Analyzing spectral pairs with anti-symmetric spectral alignment

Figure 1b illustrates case (ii) in the definition of spectral pairs. The situation becomes less transparent in case (i), namely when modification/mutation occurs in the middle of peptide (Figure 2). In this case both detecting spectral pairs (S_i, S_j) and further processing them into spectra $S^b_{i,j}$ and $S^y_{i,j}$ is more complicated.

[6] We remark that the assignments of upper indexes to spectra $S^b_{1,i}$ and $S^y_{1,i}$ are arbitrary and it is not known in advance which of these spectra represents b ions and which represents y ions.

1 KQGGTLDD LEE QAREL
2 KQGGTLDD LEE QARE
3 KQGGTLDD LEE QAR
4 KQGGTLDD LEE QA
5 KQGGTLDD LEE^{-18}QAR
6 KQGGTLDD LEE^{-18}Q
7 QGGTLDD LEE QAR
8 QGGTLDD^{+53}LEE QAR

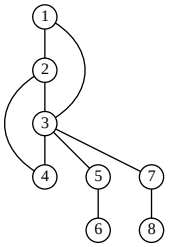

Fig. 3. A connected component of the spectral pairs graph (as projected to peptides).

Below we describe a general algorithm for deriving virtual spectra $S_{i,j}$ from spectral pairs that covers the case of internal modifications/mutations.

Let S_1 and S_2 be two spectra, and assume w.l.o.g. that $M(S_1) < M(S_2)$, where $M(S)$ denotes the parent mass of S. Let $\Delta = M(S_2) - M(S_1)$. For simplicity, we shall assume in the following that the masses in S_1 and S_2 are integers. Furthermore, we assume that S_i ($i = 1, 2$) contains the masses 0 and $M(S_i)$.

Denote by $\mathcal{M}(S_1, S_2)$ the spectral product matrix of S_1 and S_2. We define a *path* in $\mathcal{M}(S_1, S_2)$ to be a set of points in $\mathbb{R}^2$ that is composed of two diagonal segments $\{(x, x) : a \leq x < b\}$ and $\{(x, x + \Delta) : b < x \leq c\}$ for some $a \leq b \leq c$. Note that the first segment is on the blue diagonal and the second segment is on the red diagonal (one of the segments is empty when $a = b$ or $b = c$). We say that the *endpoints* of the path are the leftmost and rightmost points on the path.

The spectral alignment algorithm [16] finds the path from $(0, 0)$ to $(M(S_1), M(S_1) + \Delta)$ that contains the maximum number of points from $\mathcal{M}(S_1, S_2)$. For the optimal path P, the projection of P onto S_i (i.e. the set $\{x_1 : (x_i, x_{i-1}) \in P\}$) gives a subset of S_i which usually contains many b-ion peaks. However, this set can also contain many peaks corresponding to y and neutral loss ion peaks. In order to obtain better b/y separation, we change the spectral alignment problem by selecting only a subset of the points in P: (1) Since the minimum mass of an amino acid is 57 Da, we will choose peaks with distance at least 57 between every two peaks, and (2) We will not select two points that are generated by a peak and its complement peak in S_1 or S_2.

Formally, we say that two peaks x and x' in a spectrum S are *complements* if $x + x' = M(S) + 18$. A subset A of a spectrum S is called *anti-symmetric* if it does not contain a complement pair. A set A is called *sparse* if $|x - x'| \geq 57$ for every $x, x' \in A$. Given a path P, a set $A \subseteq P$ is called sparse if the projection of A onto S_1 is sparse, and it is called anti-symmetric if the projections of A onto S_1 and S_2 are anti-symmetric (w.r.t. S_1 and S_2, respectively). Our goal is to find the largest sparse anti-symmetric subset of $\mathcal{M}(S_1, S_2)$ that is contained in some path from $(0, 0)$ to $(M(S_1), M(S_1) + \Delta)$, and contains the points $(0, 0)$ and $(M(S_1), M(S_1) + \Delta)$.

Our algorithm for solving the problem above is similar to the algorithm of Chen et al. [4] for de-novo peptide sequencing. But unlike de-novo peptide sequencing, our problem is two-dimensional, and this adds additional complication to the algorithm. We use dynamic programing to compute optimal sets of points that are contained in two paths, one path starting at $(0, 0)$ and the other path starting at $(M(S_1), M(S_1)+\Delta)$. By keeping two paths, we make sure that for each set of points we build, its projection on S_1 is anti-symmetric. In order to keep the projection on S_2 anti-symmetric, we need additional information which is kept in a third dimension of the dynamic programming table. The full details of the algorithm are given in the appendix.

3.5 Spectral stars

A set of spectra incident to a spectrum S_1 in the spectral pairs graph is called a *spectral star*. For example, the spectral star for the spectrum derived from peptide 3 in Figure 3 consists of multiple spectra from 5 different peptides. Even for a single spectral pair (S_1, S_2), the spectra $S^b_{1,2}$ and $S^y_{1,2}$ already have high signal-to-noise ratio and rich prefix and suffix ladders. Below we show that spectral stars allow one to further enrich the prefix and suffix ladders (see Table 1). A spectral star consisting of spectral pairs (S_1, S_2), (S_1, S_3), …, (S_1, S_n) allows one to increase the signal-to-noise ratio by considering $2(n-1)$ spectra $S^b_{1,i}$ and $S^y_{1,i}$ for $2 \le i \le n$. We combine all these spectra into a *star spectrum* S^*_1 using our clustering approach. This needs to be done with caution since spectra $S^b_{1,i}$ and $S^y_{1,i}$ represent separate prefix and suffix ladders. Therefore, one of these ladders needs to be reversed to avoid mixing prefix and suffix ladders in the star spectrum. The difficulty is that the assignments of upper indexes to spectra $S^b_{1,i}$ and $S^y_{1,i}$ are arbitrary and it is not known in advance which of these spectra represents b ions and which represents y ions (i.e., it may be that $S^b_{1,i}$ represents the suffix ladder while $S^y_{1,i}$ represents the prefix ladder).

A similar problem of reversing DNA maps arises in *optical mapping* (Karp and Shamir, 2000 [11], Lee et al., 1998 [13]). It was formalized as *Binary Flip-Cut* (BFC) Problem [6] where the input is a set of n 0-1 strings (each string represents a snapshot of a DNA molecule with 1s corresponding to restriction sites). The problem is to assign a *flip* or *no-flip* state to each string so that the number of consensus sites is maximized. We found that for the case of spectral stars, a simple greedy approach to the BFC problem works well. In this approach, we arbitrarily select one of the spectra $S^b_{1,i}$ and $S^y_{1,i}$ and denote it $S_{1,i}$. We select $S_{1,2}$ as an initial consensus spectrum. For every other spectrum $S_{1,i}$ ($2 \le i \le n$), we find whether $S_{1,i}$ or its reversed copy $S^{rev}_{1,i}$ better fits the consensus spectrum. In the former case we add $S_{1,i}$ to the growing consensus, in the latter case we do it with $S^{rev}_{1,i}$.

After the greedy solution of the BFC problem we know the orientations of all spectra in the spectral star. The final step in constructing *star spectrum* S^* from the resulting collection of $S_{1,i}$ spectra using the consensus spectrum approach from Section 3.1. Table 1 illustrates the power of spectral stars in further enriching the prefix/suffix ladders.

4 Interpretation of spectral pairs/stars

The high quality of the spectra derived from spectral pairs ($S_{i,j}$) and spectral stars (S^*_i) makes *de novo* interpretation of these spectra straightforward (Figure 4). Since these spectra feature excellent separation of prefix and suffix ladders and a small number of noise peaks, de novo reconstructions of these spectra produce reliable (gapped) sequences that usually contain long correct tags.[7] On average, de novo reconstructions of our consensus spectra correctly identify 72% of all possible "cuts" in a peptide (i.e., on average, $0.72 \cdot (n-1)$ b-ions (y-ions) in a peptide of length n are explained). This is a very high number since the first (e.g., b_1) and the last (e.g., b_{n-1}) b-ions are rarely present in the MS/MS spectra thus making it nearly impossible to explain more than 80% of "cuts" in the IKKb sample. Moreover, on average, the explained b-peaks account for 95% of the total score of the de novo reconstruction implying that unexplained peaks usually have very

[7] We use the standard longest path algorithm to find the highest scoring path (and a set of suboptimal paths) in the spectrum graph of spectra $S_{i,j}$ and S^*_i. In difference from the standard de novo algorithms we do not insist on reconstructing the entire peptide and often shorten the found path by removing its prefix/suffix if it does not explain any peaks. As a result, the found path does not necessarily start/end at the beginning/end of the peptide. We also do not invoke the antisymmetric path restriction [4] since the spectra $S_{i,j}$ and S^*_i already separate prefix and suffix ladders.

low scores.[8] In addition to the optimal de novo reconstruction, we also generate suboptimal reconstructions and long peptide tags.

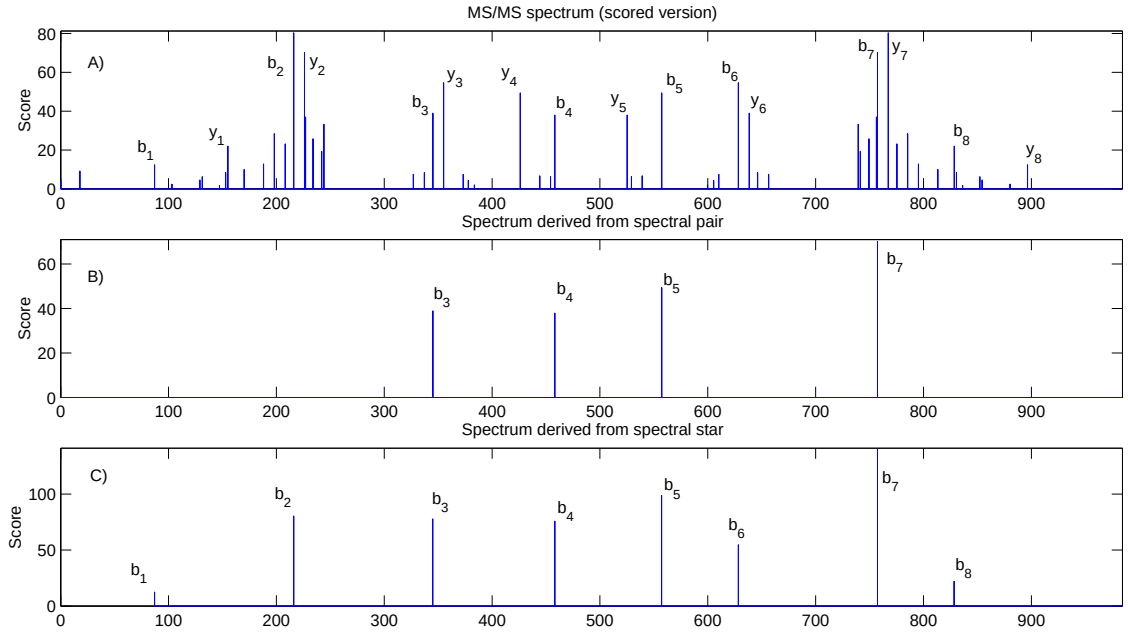

Fig. 4. Improvements in signal-to-noise. The scored MS/MS spectrum for peptide SEELVAEAH has both prefix and suffix peaks along with several noise peaks (A). Using the spectral product of a pair of spectra, many of the noise and suffix peaks that do not reside on the selected diagonal are eliminated. Though paired spectra provide very good separation of prefix/suffix ladders they may sometimes be too selective (e.g. causing the loss of the b_1, b_2, b_6, b_8 peaks) (B). By incorporating more paired spectra to form a spectral star, all noise peaks are removed and all missing prefix peaks are adequately recovered (C).

Benchmarking in mass-spectrometry is inherently difficult due to shortage of manually validated large MS/MS samples that represent "golden standards". While the ISB dataset [12] represents such a golden standard for unmodified peptides, large validated samples of spectra from modified peptides are not currently available. As a compromise, we benchmarked our algorithm using a set of 11760 spectra from the IKKb dataset that were annotated by InsPecT (with p-values$\leq$0.05) and extensively studied in recent publications [20, 21] (including comparisons with Sequest, Mascot and X!Tandem). The entire analysis (starting from clustering and ending with interpretations) of the IKKb dataset took 32 minutes on a regular desktop machine, well below the expected running time of searching the same dataset against even a medium sized database. Below we give results for both spectral pairs and spectral stars.

InsPecT identified 515 unmodified peptides[9] in the IKKb sample, 413 of which have some other prefix/suffix or modified variant in the sample and are thus amenable to pairing. We were able to find spectral

[8] We realize that our terminology may be confusing since, in reality, it is not known whether a spectrum $S^b_{i,j}$ describes b- or y-ions. Therefore, in reality we average between prefix and suffix ladders while referring to b-ions.

[9] We remark that 99 of them are represented by a single spectrum and thus are more likely to be interpretation artifacts.

pairs for 386 out of these 413 peptides. Moreover, 339 out of these 386 peptides had spectral pairs coming from two (or more) different peptides, i.e., pairs (S_1, S_2) and (S_1, S_3) such that spectra S_2 and S_3 come from different peptides.

The average number of (gapped) de novo reconstruction (explaining at least 85% of optimal score) for spectral stars was 10.4. While the spectral stars generate a small number of gapped reconstructions, these gapped sequences are not well suited for fast membership queries in the database. We therefore transform every gapped de-novo reconstruction into an ungapped reconstruction by substituting every gap with all possible combinations of amino acids.[10] On average, it results in 165 sequences of length 9.5 per spectrum. It turned out that for 86% of peptides, one of these tags is correct.

While checking the membership queries for 165 sequences[11] can be done very quickly with database indexing (at most one of these sequences is expected to be present in the database), there is no particular advantages in using such super-long tags (9.5 amino acids on average) for standard database search: a tag of length 6-7 will also typically have an unique hit in the database. However, the long 9-10 amino acid tags have distinct advantages in difficult non-standard database searches, e.g., discovery of new alternatively spliced variants or fusion genes via MS/MS analysis. Moreover, for standard search one can generate the smaller set of shorter (6-7 amino acids) tags based on the original gapped reconstruction and use them for membership queries. We used the obtained gapped reconstruction to generate such short 6 aa tags (each such tag was allowed to have at most one missing peak) and enumerated all possible continuous l-mers by substituting every gap with all possible combinations of amino acids.[12] On average, each consensus spectrum generates about 50 6-mer tags. It turned out that 82% of spectra derived from spectral stars contain at least one correct 6-mer tag.

5 Using spectral pairs to identify post-translational modifications

Our approach, for the first time, allows one to detect modifications without any reference to a database. The difference in parent masses within a spectral pair either correspond to a modification offset (case (i) above) or to a sum of amino acid masses (case (ii)). Therefore, the modification offsets present in the sample can be revealed by the parent mass differences within spectral pairs while their positions and specificities can be determined from *de-novo* reconstructions. While not every difference in parent mass corresponds to a PTM offset (some spectral pairs may be artifacts), the histogram of parent mass differences (Fig. 5) reveals the PTMs present in the IKKb sample. Indeed, 7 out of 8 most frequent parent mass differences in Fig. 5 are listed among 8 most common PTMs in IKKb sample in Tsur et al., 2005 [21]. We emphasize that Fig. 5 was obtained without any reference to database while Tsur et al., 2005 [21] found these PTMs via database search. The only modification from [21] not represented in Fig. 5 is deamidation with (small) offset 1 that is difficult to distinguish from parent mass errors and isotopic peaks artifacts. Interestingly enough, our approach reveals an offset +34 (present in thousands of spectral pairs) that was missed in [21].

[10] In rare cases the number of continuous sequences becomes too large. In such cases we limit the number of reconstructions to 500.

[11] The actual number of queries is twice as large since we have to check every "reversed" sequence as well. However, this doubling in the number of database queries can be avoided by accounting for reverse variants during the database indexing step.

[12] In rare cases the number of continuous sequences becomes too large. In such cases we limit the number of reconstructions to 100.

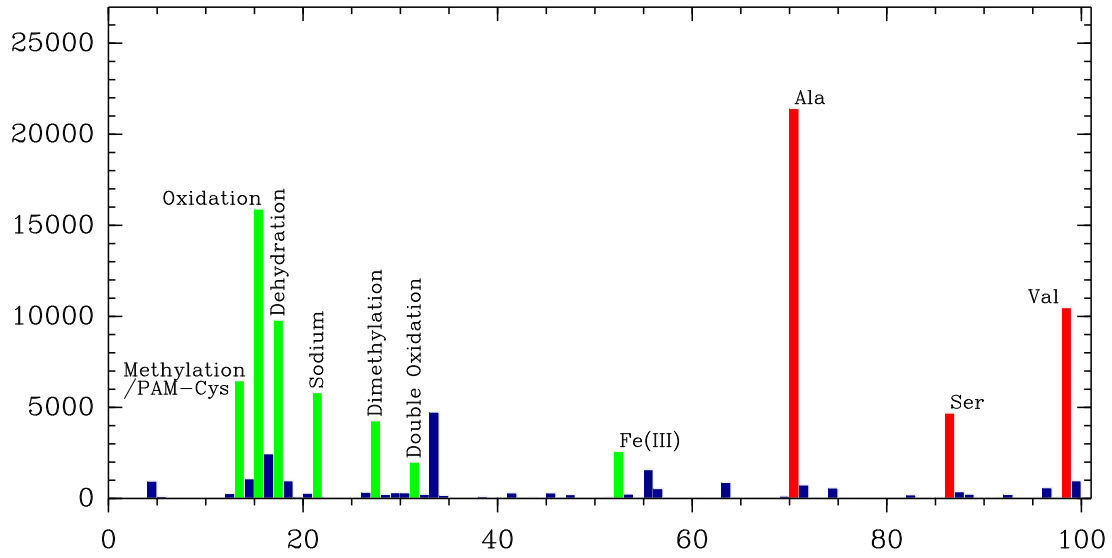

Fig. 5. Histogram of absolute parent mass differences for all detected spectral pairs; the y-axis represents the number of spectral pairs with a given difference in parent mass. For clarity, we only show mass range 1–100. The peaks at masses 71, 87, and 99 correspond to amino acid masses, and the peaks at masses 14, 16, 18, 22, 28, 32, and 53 correspond to known modifications which were also found by Tsur et al. [21]. The peak at mass 34 corresponds to a modification that remains unexplained to date.

6 Conclusions

We have demonstrated the utility of using spectral pairs and stars for protein identification. The key idea of our approach is that correlations between MS/MS spectra of modified and unmodified peptides allow one to greatly reduce noise in individual MS/MS spectra and, for the first time, make de novo interpretations so reliable that they can substitute the time-consuming matching of spectra against databases.

Tandem mass-spectra are inherently noisy and mass-spectrometrists have long been trying to reduce the noise by advancing both instrumentation and experimental protocols. In particular, Zubarev and colleagues [18, 17] recently demonstrated the power of using both CAD and ECD spectra. We emphasize that, in difference from our approach, this technique requires highly accurate Fourier transform mass-spectrometry. Another approach to reduce the complexity of spectra involves stable isotope labeling [19]. However, the impact of this approach (for peptide interpretation) has been restricted, in part by the cost of the isotope and the high mass resolution required. Alternative end-labeling chemical modification approaches have disadvantages such as low yield, complicated reaction conditions, and unpredictable changes in ionization and fragmentation. As a result, the impact of these important techniques is mainly in protein quantification rather than interpretation [19]. The key difference between our approach and labeling techniques is that, instead of trying to introduce a specific modification in a controlled fashion, we take advantage of multiple modifications naturally present in the sample. Our clustering and spectral alignment approaches allow one to decode these multiple modifications (without knowing in advance what they are) and thus provide a computational (rather than instrumentation-based or experiment-based) solution to the problem of increasing signal-to-noise ratio in MS/MS spectra.

7 Acknowledgements

The authors would like to thank Ebrahim Zandi for providing the MS/MS dataset used to benchmark our algorithm and Vineet Bafna and Stephen Tanner for insightful discussions and help in annotating the data using InsPecT. This project was supported by NIH grant NIGMS 1-R01-RR16522.

References

1. N. Bandeira, H. Tang, V. Bafna, and P.A. Pevzner. Shotgun protein sequencing by tandem mass spectra assembly. *Analytical Chemistry*, 76:7221–7233, 2004.
2. A Ben-Dor, R Shamir, and Z Yakhini. Clustering gene expression patterns. *J Comput Biol*, 6(3-4):281–297, 1999.
3. M. W. Bern and D. Goldberg. Eigenms: de novo analysis of peptide tandem mass spectra by spectral graph partitioning. *Proceedings of the 9-th annual International Conference on Research in Computational Molecular Biology (RECOMB 2005)*, pages 357–372, 2005.
4. T. Chen, M.Y. Kao, M. Tepel, J. Rush, and G.M. Church. A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *J Comput Biol*, 8(3):325–337, 2001.
5. R. Craig and R.C. Beavis. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, 20:1466 – 1467, 2004.
6. V. Dancík, S. Hannenhalli, and S. Muthukrishnan. Hardness of flip-cut problems from optical mapping. *Journal of Computational Biology*, 4(2):119–126, 1997.
7. J.K. Eng, A.L. McCormack, and J.R. Yates. An approach to correlate tandem mass-spectral data of peptides with amino acid sequences in a protein database. *Journal Of The American Society For Mass Spectrometry*, 5(11):976–989, 1994.
8. A. Frank and P. Pevzner. PepNovo: De novo peptide sequencing via probabilistic network modeling. *Analytical Chemistry*, 77:964–973, 2005.
9. A. Frank, S.W. Tanner, V. Bafna, and P.A. Pevzner. Peptide sequence tags for fast database search in mass-spectrometry. *J. of Proteome Research*, 4:1287–95, 2005.
10. E. Hunyadi-Gulyas and K. Medzihradszky. Factors that contribute to the complexity of protein digests. *Drug Discovey Today: Targets - mass spectrometry in proteomics supplement*, 3(2):3–10, 2004.
11. R. Karp and R. Shamir. Algorithms for optical mapping. *Journal of Computational Biology*, 7(1-2):303–316, 2000.
12. A. Keller, S. Purvine, A.I. Nesvizhskii, S. Stolyar, D.R. Goodlett, and E. Kolker. Experimental protein mixture for validating tandem mass spectral analysis. *OMICS*, 6(2):207–212, 2002.
13. J. K. Lee, V. Dancík, and M. S. Waterman. Estimation for restriction sites observed by optical mapping using reversible-jump Markov Chain Monte Carlo. *J. Comput. Biol.*, 5(3):505–515, 1998.
14. R.L. Levine, L. Mosoni, B.S. Berlett, and E.R. Stadtman. Methionine residues as endogenous antioxidants in proteins. *Proc Natl Acad Sci U S A.*, 93(26):15036–40, 1996.
15. D.N. Perkins, D.J. Pappin, D.M. Creasy, and J.S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, 1999.
16. P.A. Pevzner, V. Dancík, and C.L. Tang. Mutation-tolerant protein identification by mass spectrometry. *J. Comput. Biol.*, 7(6):777–787, 2000.
17. M M Savitski, M L Nielsen, F Kjeldsen, and R A Zubarev. Proteomics-grade de novo sequencing approach. *J Proteome Res*, 4(6):2348–2354, 2005.
18. M M Savitski, M L Nielsen, and R A Zubarev. New data base-independent, sequence tag-based scoring of peptide ms/ms data validates mowse scores, recovers below threshold data, singles out modified peptides, and assesses the quality of ms/ms techniques. *Mol Cell Proteomics*, 4(8):1180–1188, 2005.
19. A. Shevchenko, I. Chernushevich, W. Ens, KG. Standing, B. Thomson, M. Wilm, and M. Mann. Rapid 'de novo' peptide sequencing by a combination of nanoelectrospray, isotopic labeling and a quadrupole/time-of-flight mass spectrometer. *Rapid Commun Mass Spectrom*, 11(9):1015–1024, 1997.

20. S. Tanner, H. Shu, A. Frank, L. Wang, E. Zandi, M. Mumby, P.A. Pevzner, and V. Bafna. InsPecT: Fast and accurate identification of post-translationally modified peptides from tandem mass spectra. *Anal. Chem.*, 77(14):4626–4639, 2005.
21. D Tsur, S Tanner, E Zandi, V Bafna, and P A Pevzner. Identification of post-translational modifications by blind search of mass spectra. *Nat Biotechnol*, 23(12):1562–1567, 2005.

A The anti-symmetric spectral alignment algorithm

In this section we describe the algorithm for solving the maximum sparse anti-symmetric subset problem that was presented in Section 3.4. We use the notations and definitions from that section. For simplicity of the presentation, we will first give a simple algorithm, and then describe several enhancements to the algorithm.

Recall that the input to the problem are two PRM spectra S_1 and S_2 and the goal is to find largest sparse anti-symmetric subset of $\mathcal{M}(S_1, S_2)$ that is contained in some path from $(0,0)$ to $(M(S_1), M(S_1)+\Delta)$, and contains the points $(0,0)$ and $(M(S_1), M(S_1)+\Delta)$.

In a preprocessing stage, we remove every element x of S_1 if $x \notin S_2$ and $x+\Delta \notin S_2$. Denote $S_1 = \{x_1, \ldots, x_n\}$ and $S_2 = \{y_1, \ldots, y_m\}$, where $x_1 < x_2 < \cdots < x_n$ and $y_1 < y_2 < \cdots < y_m$. Let N be the largest index such that $x_N \leq (M(S)+18)/2$.

A peak x_i in S_1 will be called *left-critical* (resp., *right-critical*) if $x_i + \Delta \in S_1$ (resp., $x_i - \Delta \in S_1$). Denote by S_1^L and S_1^R the left-critical and right-critical peaks in S_1, respectively.

For $i \leq n$, let $\mathrm{Left}(i)$ be the set of all sparse anti-symmetric subsets of $S_1^L \cap [x_i - \Delta, x_i - 57]$, and let $\mathrm{Right}(i)$ be the set of all sparse anti-symmetric subsets of $S_1^R \cap [x_i + 57, x_i + \Delta]$. Note that if $\Delta < 57$ then $\mathrm{Left}(i) = \mathrm{Right}(i) = \phi$ for all i, which simplifies the algorithm. In the following, we shall assume that $\Delta \geq 57$.

For $i \leq N$ and $j > N$, define $D_1(i,j)$ to be the maximum size of a sparse anti-symmetric set $A \subseteq \mathcal{M}(S_1, S_2)$ such that

1. A is contained in the union of a path from $(0,0)$ to (x_i, x_i) and a path from $(x_j, x_j+\Delta)$ to $(M(S_1), M(S_1)+\Delta)$.
2. A contains the points $(0,0)$, $(M(S_1), M(S_1)+\Delta)$, (x_i, x_i), and $(x_j, x_j+\Delta)$.

If there is no set that satisfies the requirements above, $D_1(i,j) = 0$.

We define tables D_2 and D_3 in a similar way: For $i \leq N < j$ and $S \in \mathrm{Left}(i)$, $D_2(i,j,S)$ is the maximum size of a sparse anti-symmetric set $A \subseteq \mathcal{M}(S_1, S_2)$ such that

1. A is contained in the union of a path from $(0,0)$ to $(x_i, x_i+\Delta)$ and a path from $(x_j, x_j+\Delta)$ to $(M(S_1), M(S_1)+\Delta)$.
2. A contains the points $(0,0)$, $(M(S_1), M(S_1)+\Delta)$, and $(x_j, x_j+\Delta)$. Moreover, if $i > 1$ then A contains the point $(x_i, x_i+\Delta)$.
3. $\{x \in S_1^L : x_i - \Delta \leq x \leq x_i - 57 \text{ and } (x, x+\Delta) \in A\} = S$.

For $i \leq N < j$ and $S \in \mathrm{Right}(j)$, $D_3(i,j,S)$ is the maximum size of a sparse anti-symmetric set $A \subseteq \mathcal{M}(S_1, S_2)$ such that

1. A is contained in the union of a path from $(0,0)$ to (x_i, x_i) and a path from (x_j, x_j) to $(M(S_1), M(S_1)+\Delta)$.
2. A contains the points $(0,0)$, $(M(S_1), M(S_1)+\Delta)$, and (x_i, x_i). If $j < n$ then A also contains the point (x_j, x_j).

3. $\{x \in S_1^R : x_j + 57 \le x \le x_j + \Delta \text{ and } (x, x) \in A\} = S$.

We also need the following definitions: For $i \le n$, $\mathrm{prev}(i) = i'$, where i' is the maximum index such that $x_{i'} \le x_i - 57$. If no such index exists then $\mathrm{prev}(i) = 1$. Similarly, $\mathrm{next}(i) = i'$, where i' is the minimum index such that $x_{i'} \ge x_i + 57$. If no such index exists then $\mathrm{next}(i) = n$. Define

$$M_1^L(i, j) = \max_{i' \le i} D_1(i', j)$$
$$M_1^R(i, j) = \max_{j' \ge j} D_1(i, j')$$
$$M_2^R(i, j, S) = \max_{j' \ge j} D_2(i, j', S)$$

and

$$M_3^L(i, j, S) = \max_{i' \le i} D_3(i', j, S).$$

We also define

$$M_2^L(i, j, S) = \max_{i' \le i} \max_{S'} D_2(i', j, S'),$$

where the second maximum is taken over all sets $S' \in \mathrm{Left}(i')$ that are consistent with S, namely $S' \cap [x_i - \Delta, x_i - 57] = S$. Similarly,

$$M_3^L(i, j, S) = \max_{j' \ge j} \max_{S'} D_3(i, j', S'),$$

where the second maximum is taken over all sets $S' \in \mathrm{Right}(j')$ such that $S' \cap [x_j + 57, x_j + \Delta] = S$. We now show how to efficiently compute $D_1(i, j)$, $D_2(i, j, S)$, and $D_3(i, j, S)$ for all i, j, and S.

Computing $D_1(i, j)$ If either $x_i \notin S_2$ or $x_j + \Delta \notin S_2$, then by definition, $D_1(i, j) = 0$. We also have $D_1(i, j) = 0$ when x_i and x_j are complements or when $x_j - x_i < 57$. Furthermore, if $i = 1$ and $j = n$ then $D_1(i, j) = 2$. Now, suppose that none of the cases above occurs. Then,

$$D_1(i, j) = \begin{cases} M_1^L(\mathrm{prev}(i), j) + 1 & \text{if } x_i > M(S_1) + 18 - x_j \\ M_1^R(i, \mathrm{next}(j)) + 1 & \text{otherwise} \end{cases}.$$

Computing $D_2(i, j, S)$ Suppose that $x_i + \Delta, x_j + \Delta \in S_2$, x_i and x_j are not complements, and $x_j - x_i \ge 57$. If $x_{i'} + \Delta$ is complement of $x_{j'} + \Delta$ (w.r.t. S_2) for some $i' \in \{i, j\}$ and $j' \in S \cup \{j\}$, then $D_2(i, j, S) = 0$. Otherwise,

$$D_2(i, j, S) = \begin{cases} M_2^L(\mathrm{prev}(i), j, S) + 1 & \text{if } x_i > M(S_1) + 18 - x_j \\ M_2^R(i, \mathrm{next}(j), S) + 1 & \text{otherwise} \end{cases}.$$

Computing $D_3(i, j, S)$ Suppose that $x_i, x_j \in S_2$, x_i and x_j are not complements, and $x_j - x_i \ge 57$. If $x_{i'}$ is complement of $x_{j'}$ (w.r.t. S_2) for some $i' \in \{i, j\}$ and $j' \in S \cup \{j\}$, then $D_3(i, j, S) = 0$. Otherwise,

$$D_3(i, j, S) = \begin{cases} M_3^L(\mathrm{prev}(i), j, S) + 1 & \text{if } x_i > M(S_1) + 18 - x_j \\ M_3^R(i, \mathrm{next}(j), S) + 1 & \text{otherwise} \end{cases}.$$

Computing $M_1^L(i,j)$ The recurrence formula for M_1^L is straightforward: For $i = 1$, $M_1^L(i,j) = D_1(i,j)$, and for $i > 1$,

$$M_1^L(i,j) = \max\left\{D_1(i,j), M_1^L(i-1,j)\right\}.$$

The recurrence formulae of M_1^R, M_2^R, and M_3^L are similar.

Computing $M_2^L(i,j,S)$ For $i > 1$,

$$M_2^L(i,j,S) = \max\left\{D_2(i,j,S), \max_{S'} M_2^L(i-1,j,S')\right\},$$

where the second maximum is taken over all sets $S' \in \mathrm{Left}(i-1)$ that are consistent with S. The computation of $M_3^R(i,j,S)$ is similar.

Finding the optimal solution After filling the tables D_1, D_2, and D_3, we can find the size of the optimal set of points by taking the maximum value in these tables. The corresponding optimal set can be found by traversing the dynamic programming tables starting from the cell containing the maximum value.

Time complexity Using additional data structures, each cell of D_1, D_2, and D_3 can be computed in constant time (we omit the details). Thus, the time complexity of the algorithm is $O(kn^2)$, where

$$k = \max\{|\mathrm{Left}(1)|, \ldots, |\mathrm{Left}(N)|, |\mathrm{Right}(N+1)|, \ldots, |\mathrm{Right}(n)|\}.$$

Although k can be exponential in n, in practice, k has small values.

Improvements The algorithm described above can be improved in two areas. First, the accuracy can be improved by considering a variant of the maximum sparse anti-symmetric subset problem which differ from the original problem in the following aspects: (1) Each point (x, y) has a score which is equal to $score(x) + score(y)$. The goal is to find maximum weight subset that satisfies the requirement. (2) The sparse requirement is replaced by the following requirement: For every two points (x_1, y_1) and (x_2, y_2) in A, $|x_1 - x_2|$ is either greater than 200, or is equal to to the parent mass of either 1 or 2 amino acids. The algorithm described above can be easily modified to solve the new problem.

The time complexity of the algorithm can be improved by filling only part of the tables D_1, D_2, and D_3. More precisely, after some changes to the algorithm, we can fill these tables only for i and j such that $|x_i - x_j| \leq 200$. We omit the details.