A Point-Process Model for Rapid Identification of Post-Translational Modifications

Bo Yan, Tong Zhou, Peng Wang, Zhijie Liu, Vincent A. Emanuele II, Victor Olman, and Ying Xu

Pacific Symposium on Biocomputing 11:327-338(2006)

A POINT-PROCESS MODEL FOR RAPID IDENTIFICATION OF POST-TRANSLATIONAL MODIFICATIONS

BO YAN[1], TONG ZHOU[2], PENG WANG[1], ZHIJIE LIU[1], VINCENT A. EMANUELE II[2], VICTOR OLMAN[1], YING XU[1]

[1]*Department of Biochemical and Molecular Biology, University of Georgia, GA, USA*

[2]*School of Electrical and Computer Engineering, Georgia Tech, Atlanta, GA, USA*

Post-translational modifications (PTMs) are very important to biological function, and yet are notoriously difficult to detect and identify, especially in a high-throughput manner. Most of the existing approaches rely on exhaustive searches which are highly time consuming and thus are currently limited to handling of a few types of PTMs. In this paper, we present a point-process model that aims to find the optimal mass shifts to maximize the spectra alignment between an experimental MS/MS spectrum and a candidate theoretical spectrum, through cross-correlation calculation, yields a rapid search for all types of PTMs in a *blind* mode, i.e., without giving the types of the searching PTMs in advance. The test results show that our new approach's performance is comparable to or better than the other *blind* search methods, but is more efficient computationally and simpler in its concept.

1. Introduction

Post-translational modifications (PTMs) are chemical alterations in protein structures that change the properties of protein by proteolytic cleavage or by modification of amino acids [1]. They play key roles in many important cellular functions and regulatory processes. However, accurate identification of PTMs through analysis of high-throughput MS/MS data represents a highly challenging problem [2, 3]. The main difficulty lies in that the occurrences of PTMs change the molecular weights and the fragmentation patterns of peptides, which make them difficult to detect using the classical MS/MS data interpretation methods. Currently, great amount of tandem mass spectra, possibly ranging from tens of thousands to millions of mass spectra, are being collected daily across many proteomic centers and labs for functional studies of proteins. However often only a small fraction of these data could be successfully interpreted using popular analysis tools such as Sequest [4], Mascot [5], PepFrag [6] and ProteinProspector [7], etc. This can be attributed to several factors, including technical reasons such as poor peptide fragmentation, contaminants

and others. Among the biological reasons, PTMs are generally believed to be a major contributor [8].

Theoretically all possible PTMs can be identified by exhaustively searching through all types of (known) PTMs and their combinations. However, such a strategy is very time consuming; only a few types and a very small number of PTMs can be taken into account in real applications [5, 9, 10]. While recently Tanner *et al*. [11] reported a fast PTMs search method which uses peptide sequence tags [12, 13] as efficient filters to reduce the size of the database by a few orders of magnitude, this algorithm also has to "guess" the types of PTMs in advance and the time complexity depends exponentially on the number of allowed PTM types. On the other hand, most *de novo* sequencing algorithms can be modified to identify PTMs by regarding the PTMs as pseudo amino acids in additional to the 20 basic ones [14-22]. However, the requirement of relative high quality spectra such as perfect fragmentations, has seriously limited their applications.

Recently, an approach called *blind* PTM identification has been proposed by Pevzner's group [23], which allows to search for all possible types of PTMs without looking up a set of pre-specified PTMs. They reported an interesting dynamic programming approach to performing the optimal spectral alignment. The idea is that each peak segment in a theoretical spectrum is allowed to shift by one or more "appropriate" values, such that the resulting spectrum optimally matches the experimental spectrum, and from which PTMs and their locations could be derived [23]. This method could also reveal some still unknown modifications. While encouraging results have been documented, the demand for computing power to obtain the optimal spectral alignment may be too high to be practically applicable.

In this paper, we describe a point-process model [24], a time-delay estimation framework, to *blindly* search for all possible PTMs in an efficient manner. Through analyzing the cross correlation function between a query spectrum and a candidate theoretical spectrum from peptide database, as modeled by two point processes, we are able to detect all good local and global alignments between the two processes at once straightforwardly, and thus to infer the types and locations of PTMs efficiently. Spectral similarity is measured by the optimal common mass peaks shared between two spectra, short peptide segments due to missing peaks or resulting from PTMs thus can be tolerated.

We have implemented the algorithms and tested their performance on both simulated and real experimental spectra. Our approach is able to conduct *blind* PTMs search in a few seconds. For simulated spectra with 0, 1, 2 and 3 PTMs, it achieves 100%, 97%, 86% and 75% success rates respectively. The performance

on experimental spectra is comparable with or better than Tsur *et al.*'s result [23].

2. Algorithm

2.1. *A point process model*

We model a tandem mass spectrum by a point process [24]:

$$x(t) = \sum_{i=1}^{N} \delta(t - t_i)$$

where $\{t_i\}$ is a set of mass peak locations with N peaks and $\delta(t)$ is the Kronecker delta function [25]. It follows easily that

$$\delta(t - t_i)\delta(t + \tau - t_j) = \begin{cases} \delta(t - t_i) & \text{if } \tau = t_j - t_i \\ 0 & \text{otherwise} \end{cases}$$

To deal with PTMs, we assume, without loss of generality, that $\{t_i\}$ can be clustered into K+1 groups (to model K mass shifts) such that

$$x(t) = \sum_{k=0}^{K} x_k(t)$$

where each $x_k(t)$ is described by a point process model $x_k(t) = \sum_{i=1}^{N_k} \delta(t - t_i^{(k)})$, $N_k \geq 1$ and $\bigcup_{k=0}^{K} \{t_i^{(k)}\} = \{t_i\}$. N_k is the number of peaks in group k.

We further introduce a measure $C[\cdot]$ to be the total number of non-zero values in a point process, i.e., $C[x_k(t)] = N_k$ and $C[x(t)] = \sum_{k=0}^{K} N_k = N$.

When a PTM happens, a particular shift occurs to $x_k(t)$ to produce $y_k(t)$:

$$y_k(t) = x_k(t - \Delta_k) = \sum_{i=1}^{N_k} \delta(t - \Delta_k - t_i^{(k)})$$

and the resulting PTM spectrum is $y(t) = \sum_{k=0}^{K} y_k(t)$. Obviously, $C[y(t)] = N$.

Note that here we have separated the spectral peaks into K+1 different groups $\{x_k(t), y_k(t)\}$ according to their shift patterns in light of PTMs.

To simplify the analysis, we assume that all pair-wise differences among $\{t_i^{(k)}\}$ are different for all i, k and Δ_k. Then, we have

$$C[x(t - \Delta_k)y(t)] = C[y_k(t)y_k(t)] = N_k$$

Define $c_{xy}(\tau) \equiv C[x(t - \tau)y(t)] / N$, we infer that

$$c_{xy}(\tau)\,|_{\tau=\Delta_k} = \frac{N_k}{N}$$

$$c_{xy}(\tau)\,|_{\tau=(t_j^{(k)}+\Delta_k)-t_i^{(l)}} = \frac{1}{N}$$

$$c_{xy}(\tau) = 0, \text{ if } \tau \neq \Delta_k, \tau \neq (t_j^{(k)} + \Delta_k) - t_i^{(l)}$$

In practice, N_k is the number of the mass peaks shifted by one PTM, which is often much larger than 1. Therefore, the largest K+1 values of $c_{xy}(\tau)$ will be at $\tau = \{\Delta_k\}_{k=0}^{K}$, i.e., correspond to K mass shifts introduced by multiple PTMs (and their combinations), plus the local alignment between the unaffected portions of the two spectra. In other words, if a candidate peptide is a good match and no PTM exists in the query spectrum, the maximum of $c_{xy}(\tau)$ shall occur at $\tau = \Delta_0 \equiv 0$ (the rule used in Sequest [4]). On the other hand, if a candidate is the correct hit and the query peptide contains PTMs, the highest peaks of $c_{xy}(\tau)$ have a very good chance to be the right PTMs (and/or their combinations)[1]. This forms the basis of our approach to inferring PTMs.

2.2. *Implementation*

We have implemented two modes to perform *blind* PTMs search, (a) a homology search mode and (b) a strict match mode. We first calculate the cross-correlations between the query experimental spectrum and the theoretical one of each candidate from a peptide database, from which we feasibly obtain all non-zero $c_{xy}(\tau)$. Obviously those $c_{xy}(\tau)$ values contain all possible mass shifts to optimize the spectral alignment. The homology search mode reports the best hit which has the best spectral alignment with a set of optimal mass shifts. The strict match mode further requires that the hit peptide's molecular weight, or its molecular weight after modifications (if there are any PTMs), must be equal to that of the query peptide (at some tolerance). The search procedure is only applied to the candidates whose parent mass is at most Δ Da away from the query peptide (in this paper, Δ is set to 160 Da which is large enough to cover three typical PTMs).

Homology search mode: For each spectral alignment between the query spectrum and a peptide candidate, we record $c_{xy}(0)$ and the three additional largest values of $c_{xy}(\tau)$ at $\tau \neq 0$, say $c_{xy}(\Delta_1)$, $c_{xy}(\Delta_2)$ and $c_{xy}(\Delta_3)$. To find the best hits from our target database, we screen two best candidates, one with the highest value of $c_{xy}(0)$, and the other with the highest value of $c_{xy}(0)+c_{xy}(\Delta_1)$. We first check the best hit with the highest value of $c_{xy}(0)$. If this value is higher than a threshold, say 0.5, we consider there is a good match between the query and the candidate, and no PTM exists. Otherwise, either the correct peptide is not in the database or the query peptide has been modified. Then we check the best hit with the highest value of $c_{xy}(0)+c_{xy}(\Delta_1)$. If this value is

[1] We have observed in the dataset MOD1 (2620 experimental spectra with one PTM, *see* **Results**) that for 96.30%, 2.48% and 0.38% of cases, the PTMs correspond to the highest, second highest and third highest peaks of $c_{xy}(\tau)$, respectively.

higher than a threshold, say 0.7, we consider that there is a good match between the query spectrum and the candidate peptide if the candidate is modified by Δ_1, and we regard Δ_1 as the right PTM.

One might consider using $c_{xy}(0)+c_{xy}(\Delta_1)+c_{xy}(\Delta_2)$ (and so on) as the criterion. Our test results show that using $c_{xy}(0)+c_{xy}(\Delta_1)$ has a better performance, even for the cases with more than one PTM. One possible reason is that no matter how many PTMs exist, the first highest peaks of $c_{xy}(\tau)$ ($\tau \neq 0$) has the largest probability to be one of the right PTMs or of their combinations. However it may not hold for the rest highest peaks of $c_{xy}(\tau)$.

Strict match mode: This mode uses a parameter *K* to guess the number of PTMs.

For *K*=0, we implement it as a simple version of Sequest [4].

For *K*=1 (i.e., with one PTM), we report the top candidate with a Δ such that $|PW_{\exp}-PW-\Delta|\leq\varepsilon$ and $c_{xy}(0)+c_{xy}(\Delta)$ is maximized, where Δ represents the mass of a possible PTM which could be any value ranging from 0 to 160 Da in this paper. PW_{exp} and PW are the molecular weights of the query and the candidate peptides, respectively, and ε is their maximal difference allowed after modification (4 Da is used).

For *K*=2, we report the best candidate with a pair $\{\Delta_i,\Delta_j\}$ such that $|PW_{\exp}-PW-(\Delta_i+\Delta_j)|\leq\varepsilon$ and $c_{xy}(0)+c_{xy}(\Delta_i)+c_{xy}(\Delta_j)+c_{xy}(\Delta_i+\Delta_j)$ is maximized. Where Δ_i and Δ_j ($\Delta_i\neq\Delta_j\neq 0$) represent the masses of two possible PTMs, respectively.

Since the relationship $\sum_{k=0}^{K} c_{xy}(\Delta_k)=1$ always holds, the individual signal (peak) at $c_{xy}(\Delta_k)$ will diminish as *K* increases and will ultimately disappear into the background noise as *K* increases beyond certain value. We have found that strict match model is unsuitable to deal with the case with more than two PTMs.

2.3. *Determination of PTM positions*

The above calculation procedure itself does not provide information about the location of PTMs, if there are any. A tracing back procedure has been developed to locate the actual location of each predicted PTM based on the starts of peak shifts. Further details of this algorithm are omitted in this extended abstract.

2.4. *Statistical significance measurement*

We use the following z-score to measure the significance of the best hit,

$$\mathrm{z-score}=\frac{\mathrm{raw\_score}-<\mathrm{raw\_score}>}{\sigma_{\mathrm{raw\_score}}}$$

where raw_score represents a c_{xy} score , <raw_score> and $\sigma_{\text{raw\_score}}$ are the mean or standard deviation of the raw_scores derived from all peptide candidates.

We consider that a hit is significant if the hit has a high z-score. We found that in general, a correct peptide and its homologs have both high raw_score and z-score. To further classify them, we introduce a δ-score which is the difference in raw_score between the best two matches (equivalent to Δ_{cn} used in Sequest [4]). For a best hit with both high raw_score and z-score, if it has a large δ-score as well, we consider that the hit could be the correct peptide; otherwise, the hit is predicted to be a homolog of the correct peptide.

3. Results

Annotated tandem mass spectra with known PTMs are currently very limited in proteomics community. In this paper, we have tested our approach on three datasets: a large set of simulated spectra with PTMs, a set of annotated experimental spectra with added PTMs, and a small set of annotated experimental spectra with real PTMs.

3.1. *Datasets*

SIM_SET: Consisting of 4 subsets of simulated spectra, with 0, 1, 2, or 3 PTMs respectively. Each subset contains 10,000 spectra. The peptides are randomly chosen from yeast peptide database which is tryptically digested (allowed up to 2 missing cleavages). The lengths of chosen peptides are required to be at least 6, 8, 10, and 12 aa's for cases of 0, 1, 2 and 3 PTMs, respectively.

The set of simulated PTMs is acetylation of Lysine (+42), hydroxylation of Proline (+16), methylation of Aspartic acid or Glutamic acid (+14), oxidation of Methionine (+16), and phosphorylation of Serine (+80).

MOD_SET: Annotated high-quality yeast mass spectra from the Open Proteomics Database [26] (charge 2, Sequest Xcorr score ≥ 2.5). We constructed three subsets of modified spectra by adding 0, 1 or 2 PTMs selected from above PTM pool. We shift the peaks (here b, y ions only) of a spectrum to the selected modifications. If certain peaks are absent in the spectrum, we just skip these missing peaks. For the three subsets, we got 2657, 2620 and 2422 spectra with 0, 1, 2 PTMs respectively.

EXP_SET: 47 annotated high-quality spectra with real PTMs from Strader *et al.* [27]. 42 out of 47 are associated with one PTM, and most of them are oxidation of Methionine while a few are methylation of Lysine or Arginine. The 47 spectra are from 26 peptides of *R*. Palustris. We perform *blind* search against yeast peptide database mixed up with the 26 peptide sequences of *R*. Palustris.

All the experimental mass spectra were LCQ data which had a relative low mass resolution. We run a data preprocessing procedure as described in PepNovo [26] to filter tiny noise peaks and isotopic peaks. For cross-correlation calculation, we regard peak shifts in the range of $(\Delta - 0.5, \Delta + 0.5)$ as having the same nominal shift Δ, where Δ is an integer. However, our approach doesn't require that Δ is an integer.

3.2. *Search results on simulated spectra*

Table 1 shows the *blind* search results on the *SIM_SET*. Both search modes obtained a similar performance: for 0, 1, 2 and 3 PTMs, we got 100%, 97%, 85% and 72% of the spectra correctly identified, respectively. We consider a (best) hit as correct only if it matches the original peptide sequence exactly.

Table 1: Search results against the simulated spectra by homology search mode (a) and by strict match mode (b). SIM*k* refers to the sub datasets with *k* PTMs. Values at rank *i* are the percentages of the correct candidates reported at Top *i*. CPU is the average amount of time the program needs to analyze a spectrum (on a PC with a 2.8GHz CPU).

(a)

	rank 1	rank 2	rank 3	rank 4	rank 5	CPU(s)
SIM0	100%	0	0	0	0	1.516
SIM1	97.45%	1.53%	0.57%	0.21%	0.10%	1.563
SIM2	85.52%	3.83%	1.85%	1.15%	0.91%	1.467
SIM3	72.04%	7.12%	3.16%	1.85%	1.33%	1.556

(b)

	rank 1	rank 2	rank 3	rank 4	rank 5	CPU(s)
SIM0	100%	0	0	0	0	0.764
SIM1	97.90%	1.39%	0.39%	0.17%	0.03%	1.321
SIM2	85.60%	5.55%	2.10%	1.25%	0.66%	1.532

3.3. *Search results on experimental spectra*

Table 2 lists the search results against the *MOD_SET*. For spectra without PTMs, both *blind* search modes achieved 99% success rate, very close to the performance of Sequest [4] (note that the experimental spectra were annotated by Sequest). For spectra with one or two PTMs, the identification rates were relative lower. However, we found that there were certain correct candidates not reported exactly at the number one hit but within top five hits. If we regard them correct as well, then homology search mode achieved 65% and 20% success rates for spectra with one or two PTMs respectively, while strict match mode achieved 81% and 17% accuracies respectively. These results are comparable to or even better than Tsur *et al*.'s dynamic programming approach, which obtained 57.3% and 15.6% accuracies for spectra with one or two PTMs respectively [23].

Table 2: Search results against the experimental spectra by homology search mode (a) and by strict match mode (b). MOD*k* refers to the datasets with *k* PTMs. Total column is the percentage of the correct candidates reported within rank 5 in the hit list. *Search results against the dataset MOD2 by assuming peptides containing one PTM.

(a)

	rank 1	rank 2	rank 3	rank 4	rank 5	total	CPU(s)
MOD0	99.28%	0.41%	0.08%	0.04%	0	99.81%	0.823
MOD1	44.39%	9.77%	5.04%	3.05%	2.40%	64.65%	1.563
MOD2	11.44%	3.51%	2.31%	1.16%	1.32%	19.74%	1.604

(b)

	rank 1	rank 2	rank 3	rank 4	rank 5	total	CPU(s)
MOD0	99.21%	0.56%	0.04%	0.08%	0	99.89%	0.765
MOD1	60.38%	11.95%	4.31%	2.52%	2.01%	81.17%	1.467
MOD2	5.86%	3.43%	2.56%	2.52%	2.27%	16.64%	1.645
MOD2*	16.23%	5.08%	2.48%	1.65%	1.61%	27.05%	1.523

The technical reason for that strict match mode havs a much lower success rate for two PTMs may lie in that, finding a pair $\{\Delta_i, \Delta_j\}$ with maximization of $c_{xy}(0)+c_{xy}(\Delta_i)+c_{xy}(\Delta_j)+c_{xy}(\Delta_i+\Delta_j)$ might not be a good criterion to screen the correct candidate. Since not all the four signals can be observed simultaneously in some cases (depending on the positions of the two PTMs). Thus we searched MOD2 again by assuming k=1 (equivalent to maximizing $c_{xy}(0)+c_{xy}(\Delta_i+\Delta_j)$), a significant improvement was then achieved — 16% of spectra were identified correctly as top 1 and 27% within top 5.

Test results on the 47 experimental spectra with real PTMs were similar to those obtained for the experimental spectra with added PTMs. Since most of spectra have one PTM, only k=1 was searched by strict match mode. We got 27 spectra (i.e., 57.45% of spectra) identified correctly (ranked at top 1) by strict match mode and 24 spectra (i.e., 51.06%) correctly by homology search mode.

3.4. *Hits of homologs*

We have observed that some of the top hits have both relatively high z-score and raw_score. However they don't match the original peptide sequence exactly. We found that many of them are the homologs of the original peptides which have very similar sequences. For example, for the query peptide DGKYDLDFKNpESDK (where the lower case letter p indicates hydroxylation of Proline), our homology search mode reported a very similar peptide DGKYDLDFKNPNSDK with one mutation pE12PN. Table 3 lists some examples of the top one hits being the homologs of the query peptides. We estimated that ~20% of the poorly performed results by homology search mode are due to the reason of homologous proteins, which should be considered as partially correct. In Table 3, we also listed several important features of the

Table 3: Partial lists of homologs reported on the MOD1. [a] The lower case letter indicates the type and location of one PTM; [b] Change in mass; [c]Cross-correlation value at τ =0; [d] Accumulative intensities of common peaks shared between two spectra at 0 mass shift. Note that here the logarithm value of the relative intensity was used; [e] The first optimal mass shift (Δ_1) that maximizes the spectral alignment.

Query peptide		**Hits of homologs**				**Correct candidates**				
sequence[a]	PTM[b]	sequence	z-score	cxy(0)[c]	ΣI(0)[d]	cxy(0)	ΣI(0)	Δ_1[e]	cxy(Δ_1)	ΣI(Δ_1)
AIPGeYVTYALSGYVR	14	AIPGEYITYALSGYVR	8.82	0.57	3.51	0.33	1.54	14	0.37	-0.4
DGKYDLDFKNpESDK	16	DGKYDLDFKNPNSDK	8.95	0.39	2.04	0.36	1.05	16	0.43	8.38
DYIMSPVGNPEGPEKpNKK	16	DYIMSPVGNPEGPEKPNK	8.56	0.41	-7.45	0.28	-4.83	16	0.44	8.24
IINEPTAAAIAYGLdKK	14	IINEPTAAAIAYGLDK	8.32	0.40	-2.49	0.25	-2.45	14	0.47	7.99
IINEPTAAAIAYGLGAGk	42	IINEPTAAAIAYGLDK	8.51	0.40	1.70	0.41	1.23	42	0.38	13.57
KEDeEDKFDAMGNK	14	EDEEDKFDAMGNK	7.72	0.54	3.92	0.35	4.82	14	0.46	4.89
KGEQeLEGLTDTTVPK	14	GEQELEGLTDTTVPK	9.38	0.46	0.85	0.27	2.62	14	0.43	7.14
LIDLTQFPAFVTPmGK	16	LIDLTQFPAFVTPLGK	9.35	0.53	-7.62	0.33	0.86	16	0.47	5.53
LIDLTQFPAFVTpMGK	16	LIDLTQFPAFVTPLGK	8.95	0.53	-7.45	0.30	1.21	16	0.47	4.04
LIEAFNEIAEDSEQFdK	14	LIEAFNEIAEDSEQFEK	12.28	0.72	10.61	0.28	-2.81	14	0.50	9.94
LNKETTYDeIKK	14	LNKETTYDEIK	7.55	0.55	-4.28	0.41	-3.35	14	0.55	9.93
NFNDPEVQGdMK	14	NFNDPEVQADMK	9.31	0.73	12.86	0.36	2.03	14	0.45	7.88
NIVEFHsDHMK	80	NIVEFHSDHIK	8.10	0.55	0.65	0.30	3.69	80	0.50	8.99
NQAAMNpANTVFDAK	16	NQAAMNPSNTVFDAK	11.94	0.79	9.70	0.39	4.58	16	0.50	2.64
NQAAmNPANTVFDAK	16	NQAAMNPSNTVFDAK	9.10	0.61	2.42	0.39	7.88	16	0.50	-2.65
NQAAMNPsNTVFDAK	80	NQAAMNPANTVFDAK	9.51	0.61	1.46	0.43	4.37	80	0.61	2.76
RPkYFITANDVK	42	RPEYFITANDVK	13.79	0.95	11.58	0.41	1.58	1	0.59	9.99
SeVFSTYADNQPGVLIQVFEGER	14	SETFSTYADNQPGVLIQVFEGER	8.74	0.39	6.89	0.34	8.99	14	0.25	-1.83
SQIdEVVLVGGSTR	14	SQIDEIVLVGGSTR	9.79	0.69	8.09	0.35	4.14	14	0.46	-0.35
sQVDEIVLVGGSTR	80	SQIDEIVLVGGSTR	6.94	0.50	1.92	0.42	4.09	80	0.42	1.38
TAGIQIVADDLTVTNpAR	16	TAGIQIVADDLTVTNPK	10.11	0.41	1.97	0.38	-1.11	16	0.47	14.5
VATTGEWdKLTQDK	14	VATTGEWEKLTQDK	11.59	0.77	9.13	0.31	0.76	14	0.50	9.15
VdIIANDQGNR	14	VEIIANDQGNR	11.57	0.90	0.98	0.40	6.65	14	0.55	-7.65
YLdQVLDHQR	14	YLEQVLDHQR	11.65	0.94	10.74	0.39	10.83	14	0.56	-0.09

correct candidates. As we have expected that the vast majority of the inferred optimal modifications for the correct candidates correspond to the correct PTM.

4. Discussion and Conclusion

We have presented a point-process model for rapid *blind* PTMs search without the need of a list of pre-specified PTMs. Our test results show that its performance is comparable to or better than Tsur *et al*.'s dynamic programming approach [23]. Since both approaches aim to find a set of optimal mass shifts to maximize the spectral alignment, it is not surprising that they have a similar *blind* search performance. However, our algorithm is able to feasibly obtain all possible mass shifts (naturally includes the optimal mass shifts) using one round of cross-correlation calculation, thus it is conceptually simple and more computationally efficient than otheirs. Moreover, the computing time of our algorithm is independent of the number of PTMs, one major merit compared to most of the existing approaches, for which the computing time grows exponentially at the size of the set of pre-specified PTMs. We also implemented a homology search mode which is able to find the homologs of a query peptide. This feature is also found to be useful in mass spectra interpretation. Furthermore, since the similarity between two spectra is measured by the shared common mass peaks, our algorithm can tolerate short peptide segments resulted from multiple PTMs or missing peaks.

Cross correlation function has long been used to measure the similarity between two time-dependent signals. Both our approach and Sequest [4, 9] use it to measure the spectral similarity between two spectra. However our approach is significantly different from that employed in Sequest. First, Sequest performs an exhaustive search to identify PTMs for which a set of pre-specified PTMs must be given in advance. Second, Sequest considers the cross correlation value at $\tau = 0$ only. If the experimental spectrum contains PTMs, Sequest enumerates all possible PTM modifications for each candidate peptide, shifts mass peaks in a theoretical spectrum for each PTM and then compares the modified theoretical spectrum with the query spectrum. Third, Sequest doesn't remove isotopic peaks before spectral comparison. Instead, it adds artificial satellite peaks in theoretical spectrum to mimic the experimental spectrum which unnecessarily increases the computing time, and even worse it might increase the false positive identification rate. Our approach has extended the overall functionality of Sequest while maintaining its sensitivity.

In this paper, our algorithm has only considered b and y ions with +1 charge state for spectral alignment, thus it is suitable to handle MS/MS with +1 and +2

charge states, but unsuitable for spectra with +3 charge state for which the daughter ions with +1 and +2 charge states are tangled together. A simple solution could be to consider b and y ions with both +1 and +2 charge states together. However, an elaborative model should be developed to deal with this case in the future. In addition, we didn't explicitly utilize peak intensities and the number of consecutive peaks, etc. in our work. A more sophisticated scoring system which incorporates these features may further increase the success rate. Note that considering neutral mass losses doesn't improve the performance of pattern match, since contaminants generally have the same neutral mass loss patterns as well [28].

Acknowledgement

This research was supported in part by National Science Foundation (#NSF/DBI-0354771 and #NSF/ITR-IIS-0407204), by Georgia Cancer Coalition under Distinguished Cancer Clinicians & Scientists Program, and by the US Department of Energy's Genomes to Life program (http://doegenomestolife.org/) under project, "Carbon Sequestration in Synechococcus sp.: From Molecular Machines to Hierarchical Modeling".

References

1. A.A. Gooley and N.H. Packer, in *Proteome Research: New Frontiers in Functional Genomics*, M.R. Wilkins, et al., Editors. Springer-Verlag. p. 65-91 (1997).
2. M. Mann and O.N. Jensen, *Nat Biotechnol*. **21**(3):255-261 (2003).
3. O.N. Jensen, *Curr Opin Chem Biol*. **8**(1):33-41 (2004).
4. J.K. Eng, A.L. McCormack, and J.R. Yates, 3rd, *J Am Soc Mass Spectrom*. **5**(11):976-989 (1994).
5. D.N. Perkins, D.J. Pappin, D.M. Creasy, and J.S. Cottrell, *Electrophoresis*. **20**(18):3551-3567 (1999).
6. D. Fenyo, J. Qin, and B.T. Chait, *Electrophoresis*. **19**(6):998-1005 (1998).
7. K.R. Clauser, P. Baker, and A.L. Burlingame, *Anal Chem*. **71**(14):2871-2882 (1999).
8. A.I. Nesvizhskii and R. Aebersold, *Drug Discov Today*. **9**(4):173-181 (2004).
9. J.R. Yates, 3rd, J.K. Eng, and A.L. McCormack, *Anal Chem*. **67**(18):3202-3210 (1995).
10. M.R. Wilkins, E. Gasteiger, A.A. Gooley, B.R. Herbert, M.P. Molloy, P.A. Binz, K. Ou, J.C. Sanchez, A. Bairoch, K.L. Williams, and D.F. Hochstrasser, *J Mol Biol*. **289**(3):645-657 (1999).

11. S. Tanner, H. Shu, A. Frank, L.C. Wang, E. Zandi, M. Mumby, P.A. Pevzner, and V. Bafna, *Anal Chem*. **77**(14):4626-4639 (2005).
12. M. Mann and M. Wilm, *Anal Chem*. **66**(24):4390-4399 (1994).
13. D.L. Tabb, A. Saraf, and J.R. Yates, 3rd, *Anal Chem*. **75**(23):6415-6421 (2003).
14. J.A. Taylor and R.S. Johnson, *Rapid Commun Mass Spectrom*. **11**(9):1067-1075 (1997).
15. P.A. Pevzner, V. Dancik, and C.L. Tang, *J Comput Biol*. **7**(6):777-787 (2000).
16. J.A. Taylor and R.S. Johnson, *Anal Chem*. **73**(11):2594-2604 (2001).
17. P.A. Pevzner, Z. Mulyukov, V. Dancik, and C.L. Tang, *Genome Res*. **11**(2):290-299 (2001).
18. B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, and G. Lajoie, *Rapid Commun Mass Spectrom*. **17**(20):2337-2342 (2003).
19. Y. Han, B. Ma, and K. Zhang. in *Proceedings of 2004 IEEE Computational Systems Bioinformatics (CSB)*. 206-215 (2004)
20. B.C. Searle, S. Dasari, M. Turner, A.P. Reddy, D. Choi, P.A. Wilmarth, A.L. McCormack, L.L. David, and S.R. Nagalla, *Anal Chem*. **76**(8):2220-2230 (2004).
21. B. Ma, K. Zhang, and C. Liang, *Journal of Computer and System Sciences*. **70**:418-430 (2005).
22. B. Yan, Y. Qu, F. Mao, V.N. Olman, and Y. Xu, *J Comput Sci Technol*. **20**:483-490 (2005).
23. D. Tsur, S. Tanner, E. Zandi, V. Bafna, and P.A. Pevzner. in *Proceedings of 2005 IEEE Computational Systems Bioinformatics (CSB)*. 157-166 (2005)
24. D.L. Snyder and M.I. Miller, *Random point processes in time and space, 2nd edition*. 1991: Springer-Verlag.
25. A.V. Oppenheim and R.W. Schafer, *Discrete-time signal processing*. 1989: Prentice Hall.
26. J.T. Prince, M.W. Carlson, R. Wang, P. Lu, and E.M. Marcotte, *Nat Biotechnol*. **22**(4):471-472 (2004).
27. M.B. Strader, N.C. Verberkmoes, D.L. Tabb, H.M. Connelly, J.W. Barton, B.D. Bruce, D.A. Pelletier, B.H. Davison, R.L. Hettich, F.W. Larimer, and G.B. Hurst, *J Proteome Res*. **3**(5):965-978 (2004).
28. Y. Fu, Q. Yang, R. Sun, D. Li, R. Zeng, C.X. Ling, and W. Gao, *Bioinformatics*. **20**(12):1948-1954 (2004).