MCP Papers in Press. Published on March 28, 2011 as Manuscript M110.002220

Gapped Spectral Dictionaries and Their Applications for Database Searches of Tandem Mass Spectra

Kyowon Jeong[1], Sangtae Kim[2], Nuno Bandeira[2], and Pavel A. Pevzner[2]

[1] Department of Electrical and Computer Engineering, University of California, San Diego, CA.

kwj@ucsd.edu

[2] Department of Computer Science and Engineering, University of California, San Diego, CA.

{sak008, bandeira, ppevzner}@ucsd.edu

Corresponding author:

Pavel A. Pevzner
Phone: 858.822.4365
Fax: 858.534.7029
Email: ppevzner@cs.ucsd.edu

Running Title:

Gapped Spectral Dictionaries

Abbreviations

aa: amino acid

PRM: Prefix Residue Mass

PSM: Peptide Spectrum Match

DP graph: Dynamic Programming graph

PD: Path Dictionary

GPD: Gapped Path Dictionary

CGPD: Compact Gapped Path Dictionary

Summary

Generating all plausible *de novo* interpretations of a peptide tandem mass (MS/MS) spectrum (Spectral Dictionary) and quickly matching them against the database represent a recently emerged alternative approach to peptide identification. However, the sizes of the Spectral Dictionaries quickly grow with the peptide length making their generation impractical for long peptides. We introduce Gapped Spectral Dictionaries (all plausible *de novo* interpretations with gaps) that can be easily generated for any peptide length thus addressing the limitation of the Spectral Dictionary approach. We show that Gapped Spectral Dictionaries are small thus opening a possibility of using them to speed-up MS/MS searches. Our MS-GappedDictionary algorithm (based on Gapped Spectral Dictionaries) enables proteogenomics applications (like searches in the six-frame translation of the human genome) that are prohibitively time consuming with existing approaches. MS-GappedDictionary generates gapped peptides that occupy a niche between accurate but short peptide sequence tags and long but inaccurate full length peptide reconstructions. We show that, contrary to conventional wisdom, some high quality spectra do not have good peptide sequence tags and introduce gapped tags that have advantages over the conventional peptide sequence tags in MS/MS database searches.

Introduction

Most peptide identification tools are rather slow since they match every tandem mass (MS/MS) spectrum against all peptides in a database (subject to constraints on the precursor mass, the enzyme specificity, and the number of missed cleavages). A faster approach would be to generate a **full-length** *de novo* reconstruction of a spectrum and to match the resulting peptide against a database. The fundamental algorithmic advantage of the latter approach is that one can pre-process the database (e.g., by constructing its suffix tree) so that matching becomes instantaneous. The only reason why most MS/MS database search tools still use the former approach is because **full-length** *de novo* peptide sequencing remains inaccurate. Even the most advanced *de novo* peptide sequencing tools [1–3] correctly reconstruct only 30 - 45% of the *complete* peptides identified in MS/MS database searches. After decades of algorithmic developments, it seems that *de novo* peptide sequencing "hits a wall" and that accurate *full-length* peptide reconstruction is nearly impossible due to the limited *information content* of MS/MS spectra (other reasons include limited understanding of fragmentation rules, co-eluted peptides, etc.). We argue that regions with low information content should be represented as mass gaps (that represent two or more amino acids) and advocate use of gapped peptides as spectral interpretations.

Kim et al., 2009 [4] recently proposed to generate *multiple de novo* reconstructions (rather than a single one) and to match them against a database (MS-Dictionary approach). Since matching peptides against a

pre-processed database is very fast, generating thousands of reconstructions still has advantages over the traditional approaches where spectra are matched against large databases. Given an MS/MS spectrum, MS-Dictionary generates the *Spectral Dictionary* [4] that contains all plausible *de novo* reconstructions of the spectrum (i.e., with scores exceeding a given threshold) and further matches them against a database. The running time of MS-Dictionary is almost independent of the database size making it a tool of choice for peptide identification in large databases [4].

Although MS-Dictionary was proved to be useful for peptides shorter than 15 amino acids (aa), it has limitations for longer peptides with large Spectral Dictionaries. For example, the size of the Spectral Dictionary for a typical 15 aa long peptide may exceed billion of peptides making it too large for MS/MS database search. We introduce MS-GappedDictionary that generates rather small *Gapped Spectral Dictionaries* (even for long peptides) thus addressing the key limitation of the Spectral Dictionaries. Gapped Spectral Dictionary is the set of *gapped peptides* (see [5]) that are derived from the full-length peptides in the Spectral Dictionary. While the concept of a gapped peptide is not new [1, 2, 6–8], constructing dictionaries of gapped peptides that account for all plausible *de novo* interpretations was not addressed before. Gapped peptides occupy a niche between accurate but short peptide sequence tags [9] and long but inaccurate full-length peptide reconstructions. The gapped peptides are both long and accurate making them well suited for *de novo*-based MS/MS database searches. In difference from short peptide sequence tags, a gapped peptide typically has a single match in a database reducing peptide identification to a single database look-up. For a typical 20-aa long peptide, the size of the Spectral Dictionary exceeds 10^{17}, while the size of the Gapped Spectral Dictionary is only $\approx 10^4$. Moreover, we show that even smaller Gapped Spectral Dictionaries with only 20 - 100 peptides are sufficient for most applications. At the same time, gapped peptides are sufficiently long for efficient database matching. For example, for a spectrum of 15-aa long peptide, the average length[1] of gapped peptides in its Gapped Spectral Dictionary exceeds 9. For all practical purposes, (gapped) peptides of length 9 are as informative as (full-length) peptides of length 15 for matching databases (unless the database size approaches 20^9). Table 1 (a) shows the Gapped Spectral Dictionary of a spectrum of peptide LNRVSQGK shown in Figure 2 (a), consisting of 7 gapped peptides (as compared to its Spectral Dictionary consisting of 92 peptides shown in Table S1 in the Supplement). We describe an efficient algorithm for constructing the Gapped Spectral Dictionaries that also computes *coverage* of each gapped peptide, reflecting the portion of plausible *de novo* reconstructions represented by a gapped peptide (see below for the definition of coverage).

Recent proteogenomics studies highlighted the importance of MS/MS searches against the six-frame translation of genomes [10–17]. However, until recently, searches against the six-frame translations of large genomes were impractical even with the fastest MS/MS search tools, let alone with traditional tools like

[1] The total number of gaps and amino acids in the gapped peptide. For example, the length of [186]DK[246]FK is 6.

SEQUEST and Mascot. Although MS-Dictionary enabled searches in the six-frame translation of the human genome with 40X speed-up over InsPecT [4], it loses many peptide identifications (compared to InsPecT) because Spectral Dictionaries of long peptides have to be truncated (leading to truncating the correct peptides in some cases). Gapped Spectral Dictionaries remedy this shortcoming of Spectral Dictionaries and nearly double the number of identified peptides in the six-frame translation of the human genome (as compared to MS-Dictionary [4]).

Table 1 (b) illustrates how gapped peptides and their coverage can be utilized for constructing the *peptide sequence tags* [9]. Tanner et al., 2005 [18] introduced *covering sets* of tags (set of tags containing at least one correct tag) and demonstrated how such sets can greatly speed-up MS/MS database searches. However, while the sizes of covering sets may vary between spectra, Tanner et al., 2005 [18] did not describe an approach for selecting (the varying number of) tags for every spectrum and did not assign rigorous probabilities to tags. While Gapped Spectral Dictionaries can be utilized for generating (varying number of) conventional peptide sequence tags along with their probabilities, Table 1 (c) illustrates that "good" peptide sequence tags (representing all peptides in the Gapped Spectral Dictionary) may be difficult to find. We show that, contrary to conventional wisdom, some high quality spectra do not have good peptide sequence tags. We therefore advocate generating *gapped* tags representing sequences of mass gaps (like [186]LK derived from the first peptide in Table 1 (c)) and demonstrate that gapped tags improve the filtration efficiency of peptide sequence tags in tag-based MS/MS database searches.

Figure 1 illustrates different modules of MS-GappedDictionary that are described below.

Experimental Procedures

Path Dictionary Problem. Most *de novo* peptide sequencing algorithms interpret spectra by analyzing paths in *spectrum graphs* [19]. We start by discussing the problem of finding suboptimal paths in *arbitrary* graphs and later describe how it relates to finding paths in the spectrum graphs.

Let $G(V, E, score, probability)$ be a *directed acyclic graph* with vertex set V, edge set E, and functions *score* and *probability* defined on its edges (Figure 3, left panel (a)).[2] Given a path in G, the *score* of the path is defined as the sum of scores of its edges, while the *probability* of the path is defined as the product of probabilities of its edges. Given a graph G with selected vertices s (*source*) and t (*sink*), and a threshold $MinScore$, the *Path Dictionary* (denoted as $PD(G, MinScore)$) is defined as the set of all paths from s to t with scores exceeding $MinScore$ (along with their probabilities). The following Path Dictionary Problem can be solved using standard algorithms for finding suboptimal paths [20].

[2] At this point, the *score* and *probability* should be viewed as arbitrary numbers assigned to the edges. Later, we will describe what *score* and *probability* mean in the context of *de novo* peptide sequencing.

`Path Dictionary Problem.` Given a directed acyclic graph G and a threshold $MinScore$, construct $PD(G, MinScore)$.

Define the *generating function* $p(x)$ as the total probability of all paths of score x from the source s to the sink t in the graph G. The generating function can be efficiently computed as the probability of node (t, x) in the *dynamic programming graph* as described in [4, 21] (Figure 3, left). $PD(G, MinLength)$ is constructed by standard *backtracking* in the dynamic programming graph.

For the *spectrum graph* of a tandem mass spectrum [19], the Path Dictionary Problem corresponds to *de novo* peptide sequencing problem when multiple (suboptimal) *de novo* reconstructions (rather than a single one) are generated.[3] Kim et al., 2008 [21] applied the generating function approach (Figure 3, left) to analyze MS/MS spectra and further demonstrated [4] how to generate the Path Dictionary (termed *Spectral Dictionary*) that contains *all* plausible *de novo* reconstructions for a given spectrum. Each path in Path Dictionary corresponds to a full-length peptide reconstruction in the Spectral Dictionary, and $\sum_{x>MinScore} p(x)$ corresponds to the *spectral probability* (p-value) defined in [4]. To generate the Spectral Dictionaries, a spectral probability $Threshold$ is fixed and $MinScore$ is selected in such a way that the spectral probability does not exceed $Threshold$.

This Spectral Dictionary approach, while useful, is not practical for long peptides (15 amino acids and longer) with large dictionaries. We bypass this problem by solving the *Gapped Path Dictionary Problem* defined below.

Gapped Path Dictionary Problem. Let H be a subset of vertices of a graph G containing the source s and the sink t (vertices of H are called *hubs*). We remark that every path on vertices in G induces a *hub path* on vertices in H by simply retaining only vertices from H in the original path. For example, a path $s \to v_1 \to v_2 \to v_3 \to v_4 \to v_5 \to v_6 \to t$ that contains hubs s, v_2, v_3, v_5, t induces a hub path $s \to v_2 \to v_3 \to v_5 \to t$. We define the probability of a hub path as the total probability of all paths inducing this hub path. The Gapped Path Dictionary $GPD(G, H, MinScore)$ is defined as the set of all hub paths induced by the paths in $PD(G, MinScore)$ (along with their probabilities).

`Gapped Path Dictionary Problem.` Given a directed acyclic graph G, a subset of its vertices H, and a threshold $MinScore$, construct $GPD(G, H, MinScore)$.

The brute-force algorithm for constructing $GPD(G, H, MinScore)$ (by constructing $PD(G, MinScore)$ and generating all hub paths induced by the paths in $PD(G, MinScore)$) is impractical for large $PD(G, MinScore)$.

[3] In the spectrum graph of a spectrum, vertices represent all (integer) masses from 0 to parent mass of the spectrum, and vertices v and v' are connected by a directed edge (v, v') if and only if there is an amino acid with (integer) mass $(v' - v)$. The *score* of the edge (v, v') is given by the PRM score [18] of the peak represented by the vertex v', and the *probability* is given by the probability that the amino acid represented by the edge (v, v') appears in a random database (a database with identically and independently distributed amino acids with probability 1/20).

Below we describe an efficient algorithm for solving the Gapped Path Dictionary Problem that does not require the construction of $PD(G, MinScore)$.

Given hubs h and h', we define $Path(h, h')$ as the set of all paths in G between h and h' that *do not pass* through other hubs. Each path in $Path(h, h')$ is characterized by its score and probability. Let $\mathcal{X}(h, h')$ be the set of scores of all paths from $Path(h, h')$ and $Prob(h, h')$ be the total probability of all paths in $Path(h, h')$. If $Prob(h, h', x)$ is defined as the total probability of all paths of score x from the set $Path(h, h')$, then $Prob(h, h') = \sum_{x \in \mathcal{X}(h,h')} Prob(h, h', x)$.

We define the *hub graph* G_H as a multigraph on the set of vertices H (Figure 3, right). For every $x \in \mathcal{X}(h, h')$, there exists an edge between h and h' with score x and probability $Prob(h, h', x)$.[4] The *score* and the *probability* of a path in G_H is defined as the sum of scores and the product of probabilities of its edges, respectively.

As the hub paths (on vertices in H) are induced by the paths in G, $GPD(G, H, MinScore)$ is the same as $PD(G_H, MinScore)$. Therefore, the Gapped Path Dictionary Problem in G is essentially the *Path Dictionary Problem* in the hub graph G_H, and we only need to compute the scores and the probabilities of the edges in G_H to solve the Gapped Path Dictionary Problem. Below, we show how to compute $Prob(h, h', x)$ for all edges of the hub graph.

Given a hub h in the graph $G(V, E, score, probability)$, we modify the score function by assigning score $-\infty$ to all edges originating at all hubs other than h. Denote the resulting score function (parameterized by h) as $score(h)$. The family of score functions $score(h)$ for all hubs $h \in H$ can be used to compute $Prob(h, h', x)$ for all pairs of hub vertices h and h'. One can prove that computing $Prob(h, h', x)$ (for all $x \in \mathcal{X}(h, h')$) is equivalent to computing the generating function for a graph $G(V, E, score(h), probability)$ with source h and sink h'. Note that a single computation of the generating function from h to the sink t for the graph $G(V, E, score(h), probability)$ gives us $Prob(h, h', x)$ for all $h' \in H$ and all $x \in \mathcal{X}(h, h')$.

After constructing the hub graph G_H, $GPD(G, H, MinScore)$ can be constructed by computing *generating function* for the graph G_H and generating all paths with score exceeding $MinScore$. Figure 3 (right) shows an example of the Path Dictionary and the Gapped Path Dictionary.

Gapped Spectral Dictionaries. So far, we represented each path in the Gapped Path Dictionary as the sequence of *edges* (rather than *vertices*) the path traverses. Since the hub graph G_H is a multigraph (that may have multiple edges of various scores between the same vertices), there can be many paths (with different scores) with identical *vertex-sets* (Figure 3, right panel (c)). We define the *Compact Gapped Path Dictionary*, denoted by $CGPD(G, H, MinScore)$, as the set of *vertex-sets* of paths in the Gapped Path Dictionary $GPD(G, H, MinScore)$, along with their *probabilities*, where the *probability* of each vertex-set in

[4] There exists $|\mathcal{X}(h, h')|$ edges between vertices h and h' in the multigraph G_H.

$CGPD(G, H, MinScore)$ is defined as the total probability of the paths in $GPD(G, H, MinScore)$ with the same vertex-set (see Table S1 in the Supplement). The algorithm for efficient generation of Compact Gapped Path Dictionaries is described in the Supplement.

For each spectrum, we construct its spectrum graph and generate a set of hubs (prefix masses). Given a spectrum graph G and a set of hubs H, paths in G correspond to peptides while vertex-sets in G_H correspond to *gapped peptides* introduced in [5]. *Gapped Spectral Dictionary* is defined as Compact Gapped Path Dictionary of the spectrum graph.

While we described an algorithm for constructing the Gapped Spectral Dictionary for a given hub set H, it remains unclear how to select hubs. The hub selection has to achieve two conflicting goals: (i) minimize the number of selected hubs to ensure that the Gapped Spectral Dictionary is small, and (ii) maximize the average length of peptides in the Compact Gapped Spectral Dictionary to ensure that the reconstructed gapped peptides are sufficiently informative.

Therefore, the goal is to select k hubs that maximize the average number of vertices per path in the Gapped Path Dictionary (weighted by their probabilities). We select hubs as k most "popular" vertices in paths from $PD(G, MinScore)$. Such ranking of vertices of the graph G can be computed by generating *Spectral Profiles* introduced in [5].[5]

Results

Datasets. We used the previously published Shewanella, HEK, and Standard datasets to benchmark MS-GappedDictionary (see [22], [14, 23], and [24] for the details of the generation of spectra in Shewanella, HEK, and Standard datasets, respectively).

`Shewanella` dataset. To benchmark the performance of MS-GappedDictionary, we adopted the Shewanella dataset composed of 18,468 charge 2 spectra from *Shewanella oneidensis* MR-1, each representing a distinct tryptic peptide [22].[6] The spectra in this dataset were acquired on an ion trap MS (LCQ, ThermoFinnigan, San Jose, CA) using ESI and were identified with InsPecT⊕MS-GeneratingFunction [18, 21] to ensure that all Peptide Spectrum Matches (PSMs) have spectral probabilities below 10^{-9}. Note that MS-GeneratingFunction was shown to improve upon other MS/MS identification tools (InsPecT, X!Tandem, and SEQUEST/PeptideProphet [21]) and in most applications, peptide identifications with spectral probabilities above 10^{-9} are of little use since they result in high FDR.[7] The analysis below is based on Shewanella dataset unless noted otherwise.

[5] The Spectral Profiles provide a better hub selection than peak intensities and PRMs [18] (see the Supplement).
[6] While this paper focuses on doubly-charged spectra, the same generating function approach works for spectra with higher charges as shown in [25].
[7] The Supplement presents analysis of the same dataset for spectral probabilities below 10^{-10} and 10^{-11}.

`Standard` dataset. Shewanella dataset is inadequate for benchmarking the (gapped) tag generation accuracy, since the tag-based tool InsPecT was used to identify the spectra in Shewanella dataset (i.e., a correct InsPecT tag was generated for every spectrum). We obtained the dataset reported in [5] collected from the Standard Protein Mix database [24]. For this study, we considered only the charge 2 spectra generated by LTQ, where the spectra were identified by SEQUEST [26] and PeptideProphet [27] that don't use tags for identifications. We further selected PSMs with spectral probabilities below 10^{-9} and formed the dataset (denoted *Standard*) with 990 charge 2 spectra of distinct peptides.

`HEK` dataset. To benchmark MS-GappedDictionary, MS-Dictionary [4], InsPecT [18], and OMSSA [28] in MS/MS searches of huge databases, we analyzed the previously published spectral dataset from the human HEK293 cell line generated in Steve Briggs' laboratory (see [14, 23] for a detailed description of this dataset). The spectra were acquired on an LTQ linear ion trap tandem mass spectrometer.

InsPecT and OMMSA were chosen for benchmarking since they represent some of the fastest MS/MS database search tools.[8] We selected 1 million spectra from *HEK293* dataset (described in [14]) for analyzing proteogenomics applications of MS-GappedDictionary (see Supplement). Since analyzing 1 million spectra even with fast tools like InsPecT is very time consuming (estimated CPU time in the search against the 6-frame translated human genome is 9 million seconds) we further selected a single run of this dataset ($\approx$ 30,000 spectra) for benchmarking. We further processed this dataset with PepNovo+ (Release 20091029) [3] to correct charges and parent masses and limited our analysis to 14,000 charge 2 spectra (denoted *HEK* dataset). The HEK dataset was searched against the six-frame translation of the repeat-masked human genome (version GRCh37 released on March 2, 2009) using MS-GappedDictionary, MS-Dictionary, InsPecT, and OMSSA.

To generate the Gapped Spectral Dictionaries, the spectral probability threshold is set to 10^{-9} for Shewanella and Standard datasets and 10^{-11} for HEK dataset (assuming that the precursor mass is known). [9] The spectral hubs are selected based on k maximal peaks in its Spectral Profile with k varying from 20 to 40.

From Gapped Spectral Dictionaries to Pocket Dictionaries. Since multiple peptides often induce the same gapped peptide, Gapped Spectral Dictionaries are typically much smaller than Spectral Dictionaries. Figure 4 shows the sizes of Gapped Spectral Dictionaries and Spectral Dictionaries for various peptide lengths. While the size of Spectral Dictionary grows as $20^{\text{peptide length}}$, the size of the Gapped Spectral Dictionary is limited by $2^{|H|}$, where $|H|$ is the number of hubs. In practice, the size of Gapped Spectral Dictionaries is much smaller than $2^{|H|}$ for sensible values of spectral probabilities. For example, for peptides of length 20,

[8] Sequest was shown to be 60 times slower than InsPecT [4] making it impractical for large proteogenomic searches.
[9] The spectral probability thresholds vary for different datasets to maintain roughly 1% FDR (see [29] for selection of the spectral probability threshold).

the size of the Spectral Dictionary exceeds 10^{17} while the size of the Gapped Spectral Dictionary is on the order of 10^4 (for $|H| = 20$).

Figure 5 shows the distribution of the lengths of the gapped peptides that are induced by the correct peptides (*correct gapped peptides*). The high average length of the correct gapped peptides (10 - 13) indicates that Gapped Spectral Dictionaries have the potential to speed up database searches.[10] Gapped peptides are classified into *short* (with length shorter than δ) and *long* (with length equal to or longer than δ), where δ is the minimum gapped peptide length threshold. Discarding short gapped peptides results in *δ-reduced Gapped Spectral Dictionary*.

A spectrum is *δ-identifiable* if its δ-reduced Gapped Spectral Dictionary contains at least one correct gapped peptide. Figure 6 shows the identifiability of spectra in the Shewanella dataset. For $\delta = 5$, the identifiability is higher than 99% for all peptide lengths. Figure 6 illustrates that there exists a tradeoff between the identifiability and efficiency of the database search controlled by the minimum length of the gapped peptide δ (increase in δ reduces the identifiability but improves the efficiency of the database search).

After generating the δ-reduced Gapped Spectral Dictionaries, we order all gapped peptides by their *coverages*, and analyze the rank of the first correct gapped peptides in this ranked list. The *coverage* of a gapped peptide is defined as the probability of the gapped peptide divided by the total probability of the peptides in the Spectral Dictionary. Figure 7 shows that the average rank of the best ranked correct gapped peptides does not exceed 100 even for long gapped peptides ($\delta = 5, 7, 9$). In fact, only 20 - 100 gapped peptides are typically sufficient to generate a correct peptide (Figure 8). As such, it suffices to generate a small subset of the Gapped Spectral Dictionary called *Pocket Dictionary* by choosing the k best-ranked gapped peptides in the δ-reduced Gapped Spectral Dictionary (k is typically 20 - 100). Figure 9 shows the identifiability of the Pocket Dictionaries compared to the identifiability in the (full-size) δ-reduced Gapped Spectral Dictionaries.[11] Throughout the paper we generate Pocket Dictionaries of size 100 with $\delta = 5$ and 20 hubs that results in high identifiability.

While we showed how to generate the *highest-scoring* gapped peptides, generation of the *highest-probability* vertex-sets (gapped peptides) in the δ-reduced Gapped Path Dictionary is described in Supplement.

From Gapped Spectral Dictionaries to gapped tags. Once the Pocket Dictionary is generated, one still needs to match gapped peptides in the Pocket Dictionary against the protein database. The current version of MS-GappedDictionary uses *gapped tags* of length 3 (see below) instead of gapped peptides to speed-up searches in huge databases. This is conceptually similar to InsPecT search with the only difference that InsPecT uses 3-aa long peptide sequence tags while MS-GappedDictionary uses gapped tags of length 3

[10] The fraction of short gapped peptides (length less than 5) is less than 0.01 regardless of the peptide length.
[11] It turns out that selecting gapped peptides based on their coverage yields better results than selecting based on their scores (see Supplement).

for filtering the database. In Supplement we sketch a more efficient algorithm (based on matching the entire gapped peptides).

Table 1 (c) demonstrates that many gapped peptides in the Gapped Spectral Dictionary may not contain peptide sequence tags. In contrast, allowing a single gap in tags (*gapped tags*) reveals a covering set of only 6 tags of length 3: [273]LK, G[242]K, S[299]K, [250]SG, ELK, and [186]LK. In contrast with peptide sequence tags, gapped tags include both gaps and amino acid masses. Below we limit our analysis to gapped tags with gaps below 500 Da[12] and analyze gapped tags of length 3 with at most one gap (i.e., gapped tags with at least 2 amino acids). Such tags are called *proper* gapped tags. We demonstrate that the proper gapped tags have better *filtration efficiency* (defined below) than peptide sequence tags.

Some masses in a gapped peptide may represent either an amino acid or a gap because 5 amino acids (N, Q, K, R, and W with masses 114, 128, 128, 156, and 186, respectively) have *composite* masses equal to the (integer) sum of two amino acid masses.[13] For example, the composite mass 114 Da could represent either N or GG. Therefore, to generate a set of proper gapped tags, one has to decide whether a composite mass in the gapped tag corresponds to a single amino acid (see Supplement for the explanation on how it is done).

To generate the set of proper gapped tags, we select at most one proper gapped tag from each gapped peptide in the Pocket Dictionary. The greedy algorithm for selecting proper gapped tags is described in Supplement. Figure 10 compares the gapped tags generated by MS-GappedDictionary with peptide sequence tags generated by InsPecT (release 20090910). With 15 (on average) proper gapped tags generated by MS-GappedDictionary (see Table S4), the average accuracy is 94.8% while the accuracy of InsPecT tags is only 87.2% with 15 peptide sequence tags and 94.7% even with 50 tags.[14] MS-GappedDictionary constructs a table of proper gapped tags as described in the Supplement. Once the table is built, finding peptides matched to a proper gapped tag is fast, and the search space for further analysis is limited to only those matched peptides. We define the *filtration efficiency* of a peptide sequence tag/gapped tag/peptide as the ratio of the number of its matches in the random database over the database size. While the filtration efficiency of a peptide (i.e., an amino acid sequence) is $1/20^{\text{peptide length}}$ (and the filtration efficiency of amino acid is 1/20), it is easy to see that the filtration efficiency of a gap of mass m is the sum of filtration efficiencies of all amino acid sequences with mass m. It turns out that large masses typically have better filtration efficiencies than amino acids.[15] This improvement translates into a superior filtration efficiency of gapped

[12] We limit the mass of the largest gap to limit the memory requirements of MS-GappedDictionary (see Supplement).

[13] In this paper, we focus on ion-trap spectra and thus limit our analysis to integer amino acid masses. However, the generating function approach can be easily adjusted to more accurate mass measurements (see [21]).

[14] The accuracy of tag generation is defined as the percentage of cases when the set of generated tags contains a correct tag.

[15] For example, gap mass [57] (integer mass of Gly) appears in $\frac{N}{20}$ positions in a random database of size N while gap mass [400] appears in $\approx \frac{N}{121}$ positions. There are 1,102 combinations of amino acids for the gap mass [400]: 42 combinations of 3 amino acids, 664 combinations of 4 amino acids, 300 combinations of 5 amino acids, and 96

tags as compared to peptide sequence tags (compare with [31] where database searches with similar gapped tags were introduced).

For each spectrum in Standard dataset, we generated tags using MS-GappedDictionary (15 proper gapped tags per spectrum on average) and InsPecT (50 peptide sequence tags per spectrum), and measured the number of matches against the Swiss-Prot database. While InsPecT reported ≈ 2 thousand peptide sequence tag matches per spectrum on average, MS-GappedDictionary reported only ≈ 420 gapped tag matches.[16] The running time to search the Swiss-Prot database was 0.36 sec for MS-GappedDictionary (including the generation of the Gapped Spectral Dictionary and the gapped tags) and 0.51 sec for InsPecT per spectrum on a desktop machine with a 2.67-GHz Intel processor.

Database search with Gapped Spectral Dictionaries. To compare MS-GappedDictionary with other database search tools (for searches in huge databases), the HEK dataset was searched against the six-frame translation of the human genome (2.8 billion amino acid residues) using MS-GappedDictionary, MS- While traditional MS/MS database pre-processing (e.g., pre-processing by parent mass) may be more specific Dictionary (ver. 20100415) [4], InsPecT (release 20090910) [18], and OMSSA (ver. 2.1.7) [28]. The search parameters used in these searches are specified in the Supplement. We plotted the peptide level FDR curve of each tool in this search using the target-decoy database approach as described in [32]. In the case of MS-GappedDictionary, two different methods to search in the database are used: the search with gapped *tags* and the search with gapped *peptides*. We use a brute-force scanning algorithm for matching gapped peptides against the database.[17].

To measure the FDR of each tool, we first generated the reversed decoy database of the six-frame translation of the human genome. The spectra in HEK dataset were searched against both the target and decoy databases. Figure 11 shows the FDR curve of each tool and illustrates that MS-GappedDictionary significantly improves on all other tools in the number of reliably identified peptides for all levels of FDR ($\approx 30\%$ improvement in the case of 1% FDR). InsPecT is shown to improve on OMSSA and MS-Dictionary. However, MS-GappedDictionary is ≈ 20 times faster than InsPecT (0.8 sec vs 17 sec per spectrum, respectively) [18]

---

combinations of 6 amino acids. Thus, the filtration efficiency of the gap mass [400] is $42 \cdot (1/20)^3 + 664 \cdot (1/20)^4 + 300 \cdot (1/20)^5 + 96 \cdot (1/20)^6 = 0.0095$. **Figure S6 in Supplement shows the filtration efficiency of masses as compared to an amino acid.**

[16] The number of peptide matches reported by MS-GappedDictionary is only about 4 - 6 when the gapped *peptides* (not gapped tags) in the Pocket Dictionaries (with size 100) are used for the same experiment. The filtration efficiency of a gapped peptide, therefore, is 10^6 - 10^7 times better than that of gapped tags or peptide sequence tags.

[17] Searching gapped peptides against a database can be done by simply scanning each gapped peptide in the Pocket Dictionary against the database. Since a more efficient search with gapped peptides will be described elsewhere, the goal of this search with gapped peptides is to study FDR rather than to establish the running time of this primitive approach.

[18] All tools used in this benchmarking preprocess the protein database. Since preprocessing time is negligible (compared to the search time), we do not report the database preprocessing times. The running times include both

[19]. OMSSA and MS-Dictionary are also fast (1.2 sec and 0.8 sec per spectrum, respectively) but their FDRs deteriorate significantly in comparison with MS-GappedDictionary.

Figure 12 shows the length distribution of peptide identifications in the HEK dataset identified with MS-Dictionary and MS-GappedDictionary (in searches against the six-frame translation of the human genome). While Both tools identified roughly the same number of *short* peptides (length less than 14 aa), MS-GappedDictionary significantly improves on MS-Dictionary in identifying *long* peptides (14 aa and longer). This is a consequence of the fact that MS-Dictionary has to truncate the (large) spectral dictionaries of long peptides resulting in loosing many peptide identifications.

In contrast to MS-Dictionary, peptides matched to gapped peptides or gapped tags generated by MS-GappedDictionary may not belong to the Spectral Dictionary. For example, a gapped peptide AT[144]GG may match to ATSGGG (in the Spectral Dictionary) and ATGSGG (not in the Spectral Dictionary). Thus, all peptides matched by MS-GappedDictionary have to be scored to remove those that are not in the Spectral Dictionary.[20] Since the number of peptides matched by MS-GappedDictionary before scoring is typically small (Table S5), the time required for removing low-scoring peptides is negligible (less than 0.01 s per spectrum).

Discussion

Gapped peptides occupy a niche between accurate but short peptide sequence tags and long but inaccurate full-length peptide reconstructions. The gapped peptides are both long and accurate making them an ideal choice for *de novo*-based MS/MS database searches. In difference from peptide sequence tags, they typically have a few matches in a database often reducing peptide identification to a single look-up in the database. While future work will focus on efficient matching of gapped peptides against large databases, we show how gapped tags can be generated from gapped peptides to effectively filter indexed databases. Furthermore, we show how the concept of *coverage* can be instrumental for ranking sparse representations of spectral dictionaries, here limited to gapped tags and gapped peptides but conceptually generalizable to any sparse representation of all plausible peptide reconstructions. **We emphasize that every gapped pep-**

target and decoy database search times. Except OMSSA, the six-frame translation of the human genome should be divided into small sub-databases due to the memory overhead (in MS-Dictionary and MS-GappedDictionary) or unexpected errors (in InsPecT). The running time of each tool is measured by summing the search times on the sub-databases.

[19] MS-GappedDictionary filters out poor quality spectra [23] and does not generate their Gapped Spectral Dictionaries.

[20] There may be multiple peptides in the database matched to the gapped peptides or gapped tags. However, MS-GappedDictionary never accept a PSM (Peptide-Spectrum Match) without scoring the *entire spectrum* against the full length peptide using MS-GF scoring function. This additional scoring step applies to all found PSMs (gapped peptides in the Pocket Dictionary are only used to filter the database). After MS-GF scoring, MS-GappedDictionary assigns p-values (spectral probability) to each PSM.

tide search must be complemented by rigorous scoring of *all* found peptide-spectrum matches (i.e., with MS-GeneratingFunction [21] as described above) to ensure that only statistically significant PSMs are reported. MS-GappedDictionary enables proteogenomics (e.g., searches against the six-frame translation of large genomes) and metagenomics (e.g., searches against 1000+ already sequenced bacterial genomes) analysis that is prohibitively slow for traditional MS/MS database search tools.

While this paper focuses on non-modified gapped peptides (proteogenomics studies are typically based on non-modified peptides, [21] MS-GappedDictionary is applicable to spectra of modified peptides as well (see Supplement). If the set of modifications is given in advance (like in traditional MS/MS search approaches), one can generate the set of modified gapped peptides by simply extending the set of masses to accommodate masses of modified amino acids. Nevertheless, the probability that the Pocket Dictionary contains a correct gapped peptide may start decreasing if diverse modifications are added to the analysis. Moreover, gapped peptides with modifications should be converted into those without modification when they are used for the database search. The algorithms that address these issues are under development.

While MS-GappedDictionary has a potential to speed-up database searches by orders of magnitudes as compared to other widely used tools such as SEQUEST and InsPecT, its performance deteriorates in the case of highly charged spectra (charge 4 and higher). This is a bottleneck for all MS/MS database search approaches based on full length peptides or peptide sequence tags [18]. Further advances in design of scoring functions for highly charged spectra are needed to address this bottleneck [25].

We emphasize that the benefits of a pre-processed database are best utilized when the database does not need to be re-processed to reflect changes in enzyme specificity, number of missed cleavages, etc. Our approach assumes a standard combinatorial pattern matching (CPM) database pre-processing (e.g., hash tables, keyword trees, suffix trees, etc. [33]) rather than a specialized MS/MS database pre-processing that may account for different search parameters such as the precursor mass or the enzyme specificity. Thus, we assume that applications of MS-GappedDictionary do not require database re-processing when the search parameters change. While traditional MS/MS database pre-processing (e.g., by parent mass) may be more specific than a CPM pre-processing, this benefit is being offset by the universal nature of CPM pre-processing and by the fact that gapped peptide searches are much faster than the traditional database searchs (even with universal rather than specialized database

[21] We remark that many modified peptides identified in typical MS/MS searches are also identified as non-modified peptides. For example, while oxidation of Met is very common, as observed in Gupta et al. 2007 [22], for a great majority of identified peptides with Met^{+16}, there exists also a non-modified version of the same peptide (that is sufficient for proteogenomics applications). This observation applies to most chemical adducts and even some biological modifications.

indexing). In the case when the search changes to include an additional post-translational modification, we suggest to change the gapped peptide generation (i.e., to transform gapped peptides with modifications into gapped peptides without modifications) rather than to re-process the database.

Acknowledgments

We are grateful to Dick Smith and Steve Briggs for making their spectral datasets available. The research was supported by the National Center for Research Resources of NIH via grant P-41-RR24851.

References

1. Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., and Lajoie, G. (2003) PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **17**, 2337-2342.
2. Frank, A., and Pevzner, P. (2005) PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal. Chem.* **77**, 964 -973.
3. Frank, A. (2009) A ranking-based Scoring Function for peptide-spectrum matches. *J. Proteome Res.,* **8**, 2241-2252.
4. Kim, S., Gupta, N., Bandeira, N., Pevzner, P. (2009) Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra. *Mol. Cell. Proteomics* **8**, 53-69.
5. Kim, S., Bandeira, N., Pevzner, P. (2009) Spectral Profiles, a Novel Representation of Tandem Mass Spectra and Their Applications for de Novo Peptide Sequencing and Identification. *Mol. Cell. Proteomics* **8**, 1391-1400.
6. Searle, B., Dasari, S., Wilmarth, P., Turner, M., Reddy, A., David, L., and Nagalla, S. (2005) Identification of Protein Modifications Using MS/MS de Novo Sequencing and the OpenSea Alignment Algorithm. *J. Proteome Res.,* **4(2)**, 546-554.
7. Johnson, S., and Taylor, A. (2002) Searching sequence databases via de novo peptide sequencing by tandem mass spectrometry. *Mol. Biotechnology* **22(3)**, 301-315.
8. Huang, L., Jacob, R. J., Pegg, S. C., Baldwin, M. A., Wang, C. C., Burlingame, A. L., and Babbitt, P. C. (2001) Functional assignment of the 20 S proteasome from Trypanosoma brucei using mass spectrometry and new bioinformatics approaches. *J. Biol. Chem.* **276**, 28327-28339.
9. Mann, M., and Wilm, M. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **66**, 4390-4399.
10. Yates, J., Eng, J., and McCormack, A. (1995) Mining genomes: Correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal. Chem.* **67**, 3202-3210.
11. Kuster, B., Mortensen, P., Andersen, J.S., and Mann, M. (2001) Mass spectrometry allows direct identification of proteins in large genomes. *Proteomics.* **1**, 641-650.
12. Choudhary, J.S., Blackstock, W.P., Creasy, D.M., Cottrell, J.S. (2001) Interrogating the human genome using uninterpreted mass spectrometry data. *Proteomics.* **1(5)**, 651-67.
13. Oshiro, G., Wodicka, L., Washburn, M., Yates III, J., Lockhart, D., and Winzeler, E. (2002) Parallel identification of new genes in Saccharomyces cerevisiae. *Genome Res.* **12**, 1210-1220.
14. Tanner, S., Shen, Z., Ng, J., Florea, L., Guigo, R., Briggs, S., and Bafna, V. (2007) Improving gene annotation using peptide mass spectrometry. *Genome Res.* **17**, 231-239.
15. Castellana, N.E., Payne, S.H., Shen, Z., Stanke, M., Bafna, V., and Briggs, S.P. (2008). Discovery and revision of Arabidopsis genes by proteogenomics. *Proc. Natl. Acad. Sci. USA 105*, 21034-21038.

16. Borchert, N., Dieterich, C., Krug, K., Schutz, W., Jung, S., Nordheim, A., Sommer, R. J., and Macek, B. (2010) Proteogenomics of Pristionchus pacificus reveals distinct proteome structure of nematode models. *Genome Research*, doi: 10.1101/gr.103119.109.

17. Baerenfaller, K., Grossmann, J., Grobei, M.A., Hull, R., Hirsch-Hoffmann, M., Yalovsky, S., Zimmermann, P., Grossniklaus, U., Gruissem, W., and Baginsky, S. (2008) Genome-Scale Proteomics Reveals Arabidopsis thaliana Gene Models and Proteome Dynamics. *Science* **320**, 938-941.

18. Tanner, S., Shu, H., Frank, A., Wang, L., Zandi, E., Mumby, M., Pevzner, P. A., and Bafna, V. (2005) InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.* **77**, 4626-4639.

19. Dancik, V., Addona, T., Clauser, K., Vath, J., and Pevzner, P. (1999) De novo protein sequencing via tandem mass-spectrometry. *J. Comp. Biol.* **6**, 327-341.

20. David Eppstein. (1998) Finding the k Shortest Paths. *SIAM J. Comput.* **28**, 652-673.

21. Kim, S., Gupta, N., and Pevzner, P. A. (2008) Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J. Proteome Res.* **7**, 3354-3363.

22. Gupta, N., Tan ner, S., Jaitly, N., Adkins, J., Lipton, M., Edwards, R., Romine, M., Osterman, A., Bafna, V., Smith, R., Pevzner, P. (2007) Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. *Genome Res.*, **17**, 1362-1377.

23. Frank, A., Bandeira, N., Shen, Z., Tanner, S., Briggs, S., Smith, R., and Pevzner, P. (2008) Clustering millions of tandem mass spectra. *J. Proteome Res.* **7**, 113-122.

24. Klimek, J., Eddes, J. S., Hohmann, L., Jackson, J., Peterson, A., Letarte, S., Gafken, P. R., Katz, J. E., Mallick, P., Lee, H., Schmidt, A., Ossola, R., Eng, J. K., Aebersold, R., and Martin, D. B. (2008) The standard protein mix database: a diverse datasetto assist in the production of improved peptide and protein identification software tools. *J. Proteome Res.* **7**, 96-103.

25. Kim, S., Mischerikow, N., Bandeira, N., Mohammed, S., Heck, A., and Pevzner, P. (2010) The generating function of CID, ETD and CID/ETD pairs of tandem mass spectra: applications to database search (submitted)

26. Eng, J., McCormack, A., and Yates, J. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **7**, 655-667

27. Keller, A., Nesvizhskii, A., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74**, 5383-5392

28. Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W., and Bryant, S. H. (2004) Open mass spectrometry search algorithm. *J. Proteome Res.* **3**, 958-964

29. Gupta, N. and and Pevzner, P. (2009) False discovery rates of protein identifications: a strike against the two-peptide rule. *J Proteome Res.* **8** 4173-4181.

30. Gupta, N., Benhamida, J., Bhargava, D., Goodman, E., Kain, I., Nguyen, N., Ollikainen, N., Rodriguez, J., Wang, J., Lipton, M., Romine, M., Bafna, V., Smith, R., and Pevzner, P. (2008) Comparative proteogenomics: combining mass spectrometry and comparative genomics to analyze multiple genomes. *Genome Res.* **18**, 1133-1142.

31. Bern, M., Cai, Y., and Goldberg, D. (2007) Lookup Peaks: A Hybrid of de Novo Sequencing and Database Search for Protein Identification by Tandem Mass Spectrometry. *Anal. Chem.* **79 (4)**, 1393-1400.

32. Elias, J.E. and Gygi, S. P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods* **4**, 207-214.

33. Gusfield,D. (1997) Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology. *Cambridge University Press, New York.*

Figure Legends

Fig. 1. Different modules of MS-GappedDictionary.

Fig. 2. Spectra for the peptide LNRVSQGK (a) and AIIDAIVSGELK (b) identified by InsPecT (release 20090910) database search.

Fig. 3. Left panel : Illustration of the dynamic programming algorithm for computing the generating function of graph G shown in (a). The nodes of the *dynamic programming (DP) graph* (b) are defined as pairs (v, x), where v is a vertex of G and x is a score. Two nodes (v, x) and (v', x') are connected by an edge if and only if there exists an edge between vertices v and v' in G with score $x' - x$. The probability of an edge between (v, x) and (v', x') in the DP graph equals to the probability of the edge (v, v') in G. A source s in graph G corresponds to a single node $(s, 0)$ in the DP graph. A node (v, x) is present in the DP graph if and only if there exist a path from $(s, 0)$ to (v, x). In this example, red (blue) edges of the DP graph in (b) are from the red (blue) edges of the graph G in (a). All edge probabilities in (b) are 0.5 as the probabilities of edges of G are 0.5. The *node probability* of node (v, x) (shown inside nodes in (b) and (c)) is the total probability of the paths from the source s to v with the score x. The node probability of the source of the DP graph is initialized by 1, and the node probability of a node (v, x) is obtained by the *weighted* summation of the node probabilities of its *predecessors* (see [21]). The generating function is represented by the probabilities of the sink nodes in the DP graph. To find all paths of score x from the source to the sink in graph G one has to backtrack all paths from the node (t, x) in the DP graph. For example, if $x = 2$, two such paths are found: $\{s, v_2, v_4, v_7, t\}$ and $\{s, v_3, v_6, t\}$ as in (c).

Right panel : Path Dictionary and Gapped Path Dictionary. (a) $PD(G, 1)$ and the generating function of G. (b) The construction of G_H using edges between hubs v_2 and t (shown as solid blue and red edges) as examples. Solid blue and red edges in G_H are induced by dashed blue and red paths in G. All paths that use only non-hub vertices in G are collapsed into edges in G_H. (c) The hub graph G_H, $GPD(G, H, 1)$, and the generating function of G_H.

Fig. 4. Gapped Spectral Dictionary size vs. Spectral Dictionary size (for varying peptide length and number of hubs) for the Shewanella dataset.

Fig. 5. Distribution of the lengths of the gapped peptides induced by correct peptides (for 20 hubs) for the Shewanella dataset. (see Supplement for different parameters).

Fig. 6. Identifiability of the δ-reduced Gapped Spectral Dictionaries from the Shewanella dataset for $\delta = 5$ (a), $\delta = 7$ (b), and $\delta = 9$ (c).

Fig. 7. Average rank of (the best ranked) correct gapped peptides. The average ranking does not exceed 80 regardless of the peptide length (for $\delta = 5, 7, 9$). The number of hubs is 20. The dotted lines with open circles at the ends represent the range that the rankings fall into 90% of the time.

Fig. 8. The probability that a correct gapped peptide is found within k top-ranked peptides in the δ-reduced Gapped Spectral Dictionary. The number of hubs is 20, and $\delta = 5$ (see Supplement for different parameters).

Fig. 9. Identifiability of the Pocket Dictionaries from the Shewanella dataset for $\delta = 5$ (a), $\delta = 7$ (b), and $\delta = 9$ (c). The number of hubs is 20. Even for long peptides, Pocket Dictionaries with 50 gapped peptides are sufficient to ensure the identifiability higher than 97% when δ is 5. When δ is large, larger Pocket Dictionaries are needed.

Fig. 10. Comparison of gapped tags generated from the Pocket Dictionaries and the peptide sequence tags generated by InsPecT (on spectra from the Standard dataset).

Fig. 11. The FDR curves for MS-GappedDictionary (using either gapped tag or gapped peptides), OMSSA, InsPecT, and MS-Dictionary (peptide-level FDR is reported [32]). For each spectrum, only the single best matching peptide is reported.

Fig. 12. The length distribution of peptides with the spectral probability less than 10^{-13} (corresponding FDR ≈1%) in HEK dataset identified by MS-GappedDictionary and MS-Dictionary in the six-frame translation of the human genome. MS-Dictionary identifies less peptides than MS-GappedDictionary when the peptide length is longer than 13.

Table. 1. (a) The Gapped Spectral Dictionary for the spectrum of peptide LNRVSQGK (consisting of 7 gapped peptides) is much smaller than the Spectral Dictionary (consisting of 92 full-length peptides). For simplicity, LNRVSQGK is represented by its *integer* amino acid masses as follows: [113][114][156][99][87][128][57][128]. Each gapped peptide is represented by amino acids and *mass gaps* that represent combinations of amino acids (for example, [128] can be Q, K, GA, or AG). Either Q or K is used instead of [128] when [128] occupies the same position as Q or K on the peptide LNRVSQGK. The gapped peptides that match the correct peptide are called *correct* gapped peptides (like gapped peptides 1 and 6 marked with †). For example, the gapped peptides [113 + 114]RVSQGK or LN[156+99]SQGK match peptide LNRVSQGK. The second column represents the coverage of the gapped peptide (see Results section for the definition of coverage), reflecting the portion of the total probability of all full-length peptides represented by the gapped peptide (see Supplement for an example of the calculation of the coverage of the gapped peptide [227]RVSQGK).
(b) Peptide sequence tags of length 3 derived from the Gapped Spectral Dictionary. Masses over left (right) arrows are the prefix (suffix) masses of the tags. The third column shows the coverage of each tag, where the coverage of a tag is defined by the summation of the coverages of gapped peptides covered by the tag. The

fourth column shows the gapped peptides (specified by the numbers in the first column of (a)) covered by each tag. For example, a tag VRV covers two gapped peptides 3 and 5 in (a) with coverages of 13.71% and 5.71%, respectively. The coverage of the tag VRV is, thus, $13.71 + 5.71 \approx 19.4\%$. Overall, only 2 tags (e.g., QGK and VRV) cover all gapped peptides in the Gapped Spectral Dictionary.

(c) The Gapped Spectral Dictionary for the spectrum of peptide AIIDAIVSGELK shown in Figure 2 (b) (16 gapped peptides represent 24,034 full length peptides). The correct gapped peptides are marked by †. The Gapped Spectral Dictionary for the peptide AIIDAIVSGELK reveals only 3 tags (GEL, ELK, and SGE), together covering only 18.59% of the Spectral Dictionary. In contrast, 6 (*gapped tags*) [273]LK, G[242]K, S[299]K, [250]SG, ELK, and [186]LK cover the entire Spectral Dictionary.

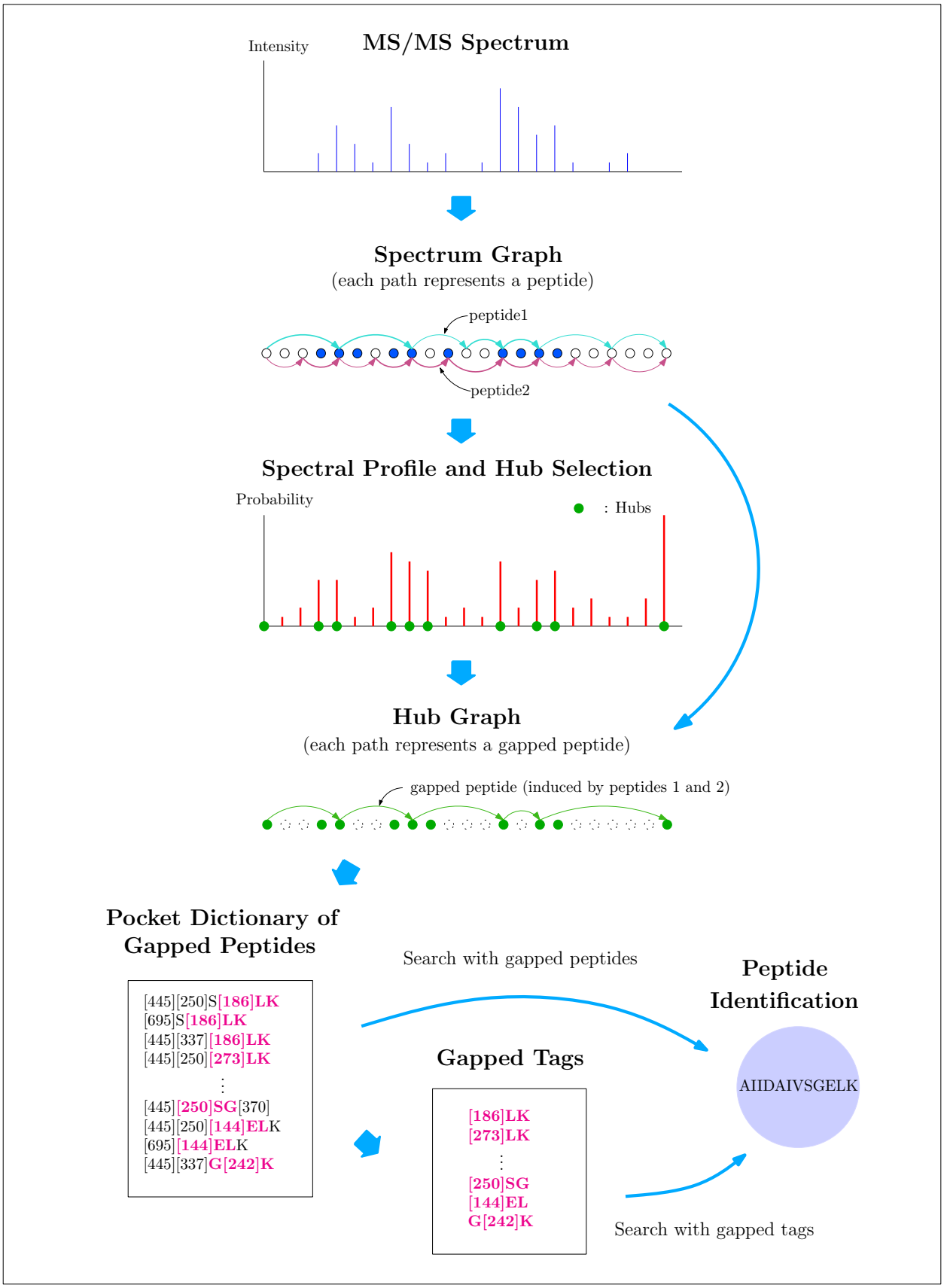
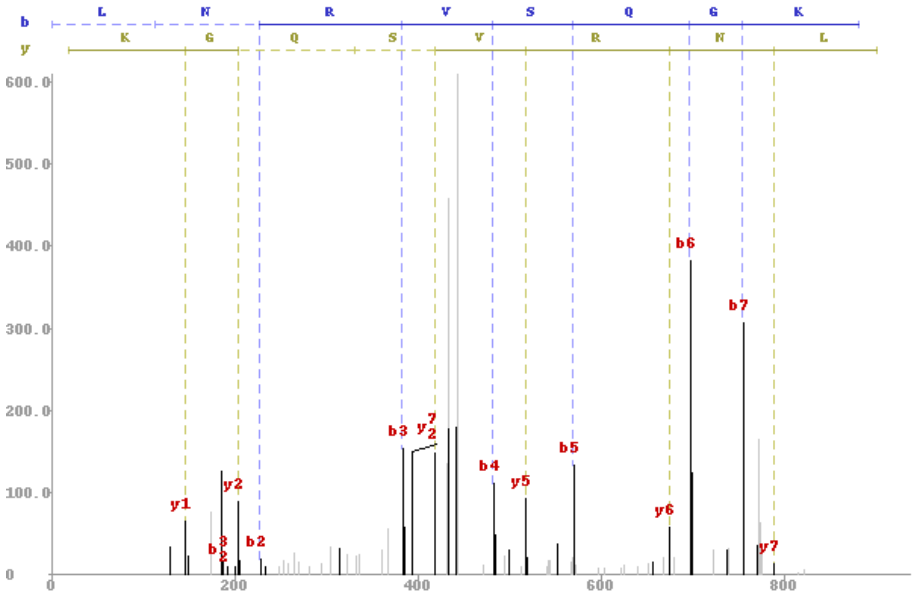
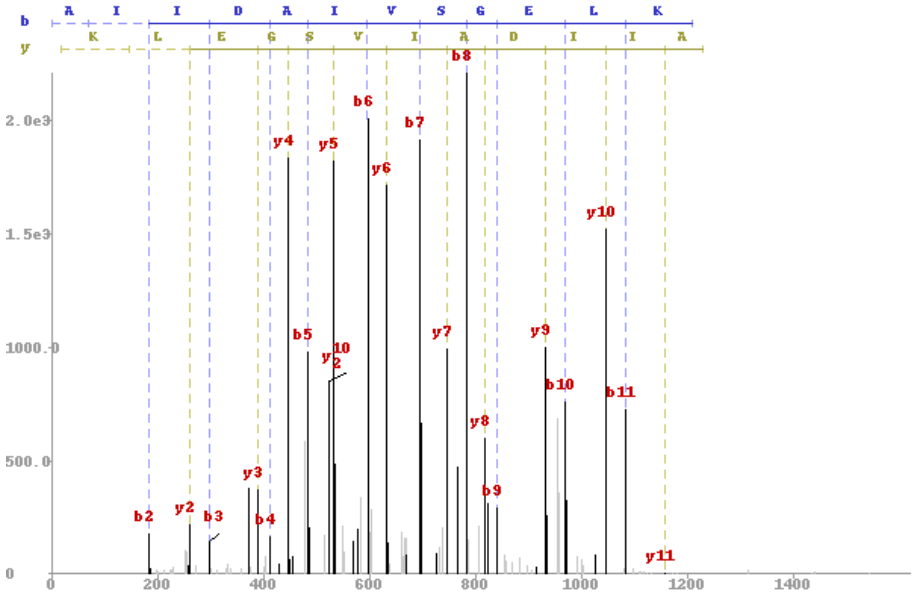

Intensity
MS/MS Spectrum
Spectrum Graph
(each path represents a peptide)
peptide1
peptide2
Spectral Profile and Hub Selection
Probability
: Hubs
Hub Graph
(each path represents a gapped peptide)
gapped peptide (induced by peptides 1 and 2)
Pocket Dictionary of Gapped Peptides
[445][250]S[186]LK
[695]S[186]LK
[445][337][186]LK
[445][250][273]LK
[445][250]SG[370]
[445][250][144]ELK
[695][144]ELK
[445][337]G[242]K
Search with gapped peptides
Gapped Tags
[186]LK
[273]LK
[250]SG
[144]EL
G[242]K
Search with gapped tags
Peptide Identification
AIIDAIVSGELK

Fig. 1.

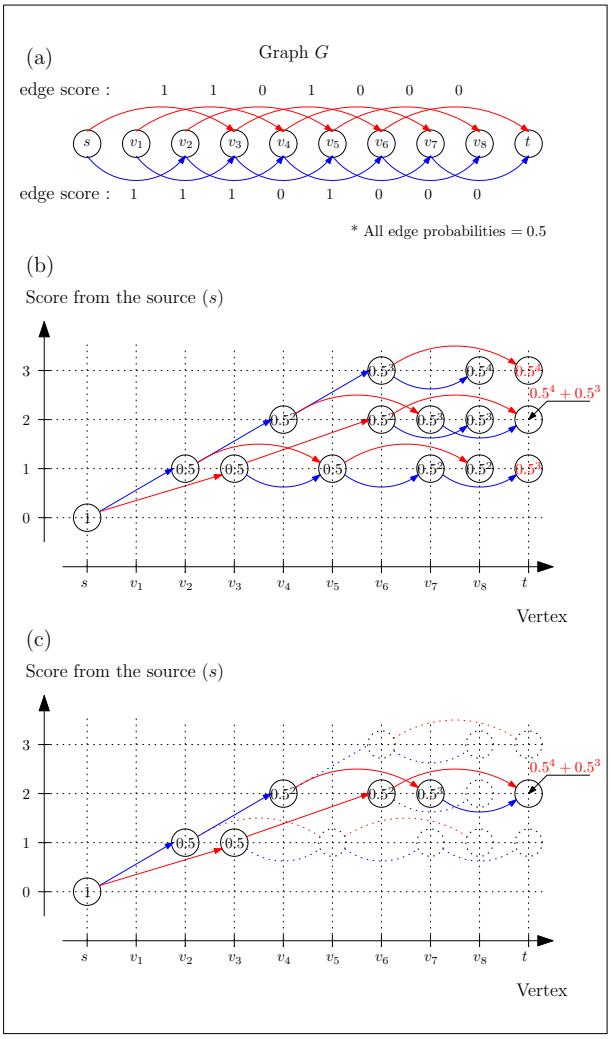


(a)



(b)

Fig. 2.


(a)
Graph G
edge score : 1 1 0 1 0 0 0
edge score : 1 1 1 0 1 0 0 0
* All edge probabilities = 0.5
(b)
Score from the source (s)
Vertex
(c)
Score from the source (s)
Vertex

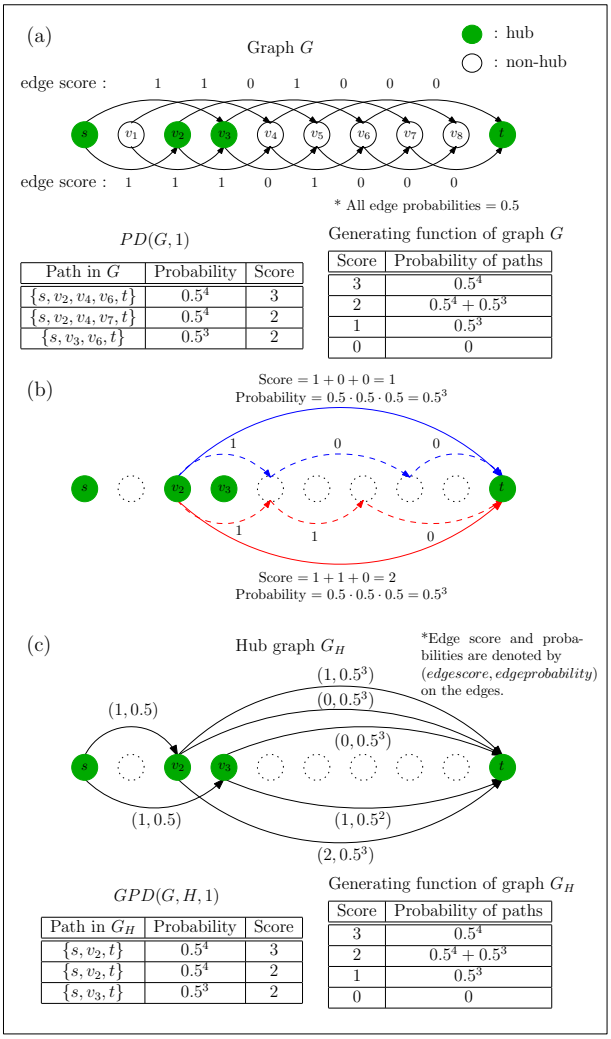

(a)
Graph G
: hub
: non-hub
edge score : 1 1 0 1 0 0 0
edge score : 1 1 1 0 1 0 0 0
* All edge probabilities = 0.5
$PD(G, 1)$
Path in G | Probability | Score
$\{s, v_2, v_4, v_6, t\}$ | 0.5^4 | 3
$\{s, v_2, v_4, v_7, t\}$ | 0.5^4 | 2
$\{s, v_3, v_6, t\}$ | 0.5^3 | 2
Generating function of graph G
Score | Probability of paths
3 | 0.5^4
2 | $0.5^4 + 0.5^3$
1 | 0.5^3
0 | 0
(b)
Score $= 1 + 0 + 0 = 1$
Probability $= 0.5 \cdot 0.5 \cdot 0.5 = 0.5^3$
Score $= 1 + 1 + 0 = 2$
Probability $= 0.5 \cdot 0.5 \cdot 0.5 = 0.5^3$
(c)
Hub graph G_H
*Edge score and probabilities are denoted by $(edgescore, edgeprobability)$ on the edges.
$GPD(G, H, 1)$
Path in G_H | Probability | Score
$\{s, v_2, t\}$ | 0.5^4 | 3
$\{s, v_2, t\}$ | 0.5^4 | 2
$\{s, v_3, t\}$ | 0.5^3 | 2
Generating function of graph G_H
Score | Probability of paths
3 | 0.5^4
2 | $0.5^4 + 0.5^3$
1 | 0.5^3
0 | 0

Fig. 3.

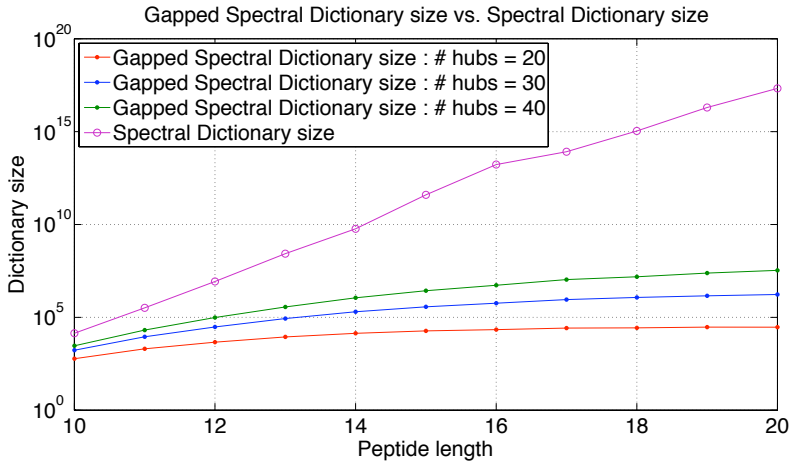

Gapped Spectral Dictionary size vs. Spectral Dictionary size
Gapped Spectral Dictionary size : # hubs = 20
Gapped Spectral Dictionary size : # hubs = 30
Gapped Spectral Dictionary size : # hubs = 40
Spectral Dictionary size
Dictionary size
Peptide length

Fig. 4.

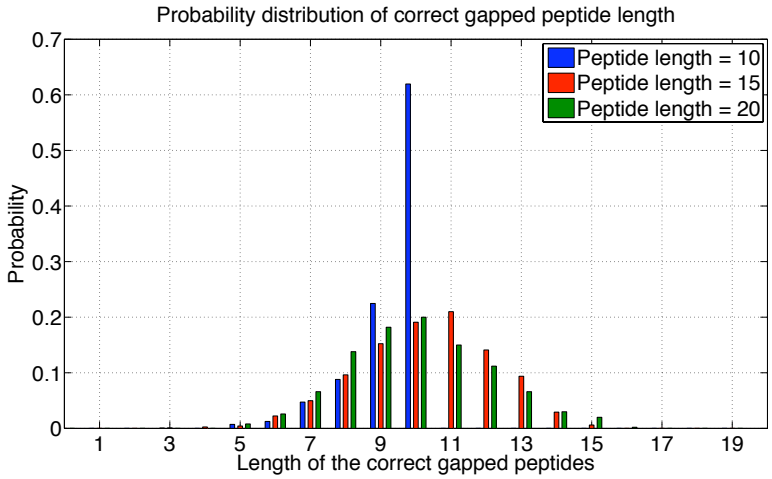
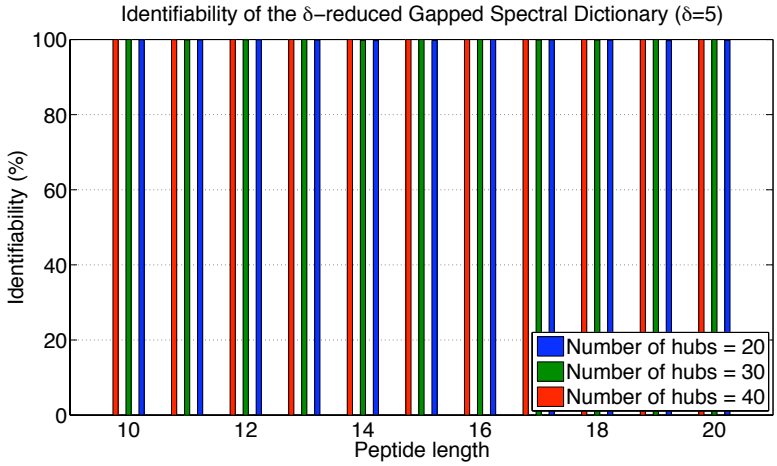
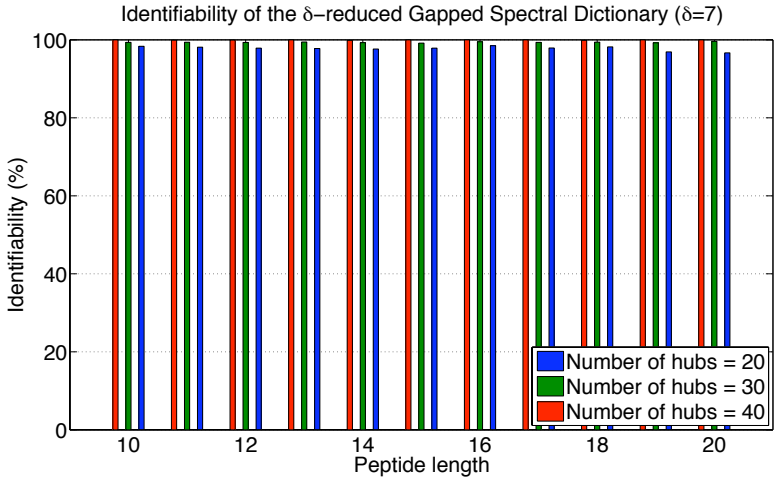


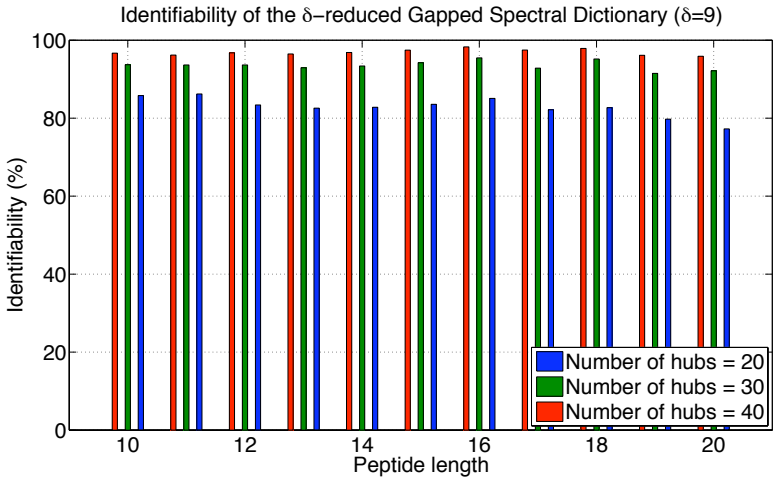
Fig. 5.



(a)



(b)



(c)

Fig. 6.

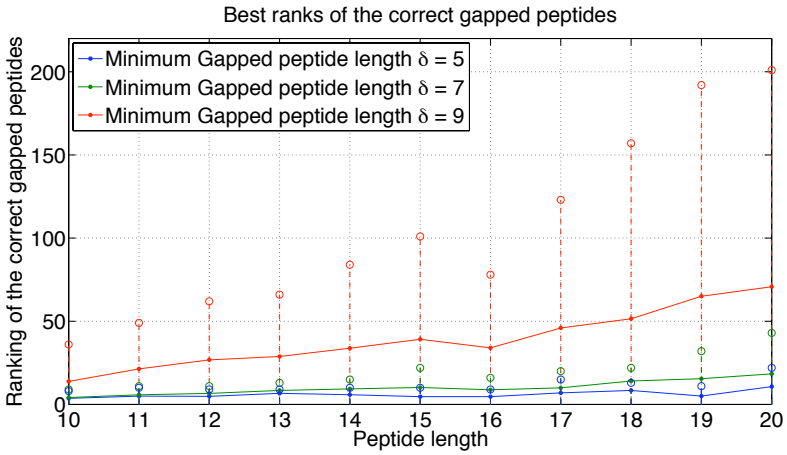


Fig. 7.

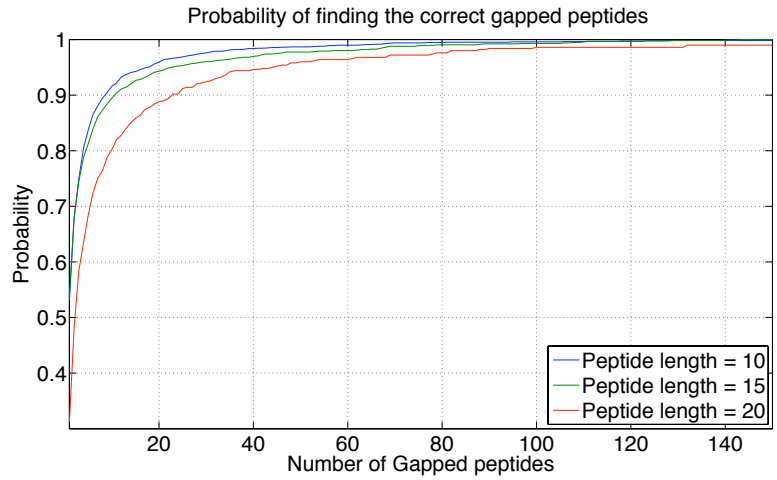
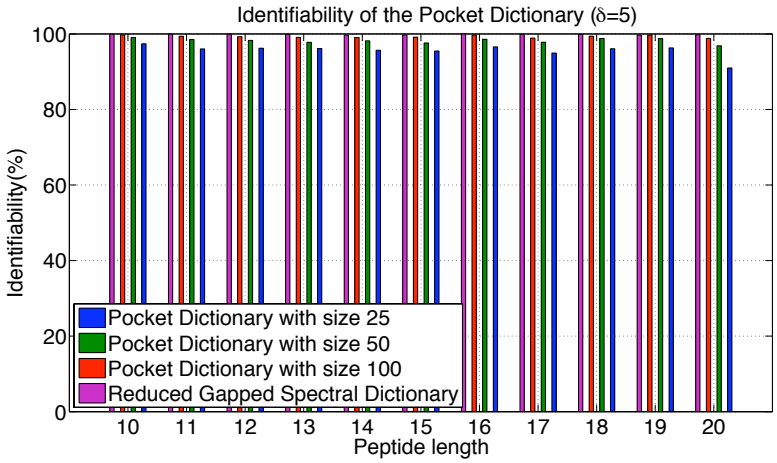
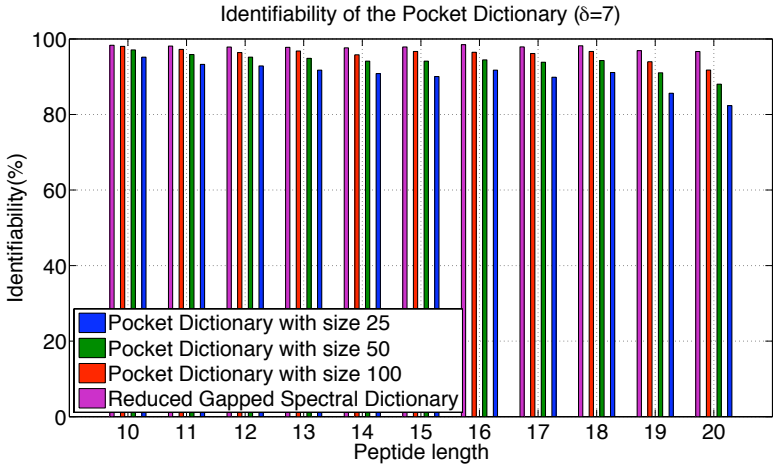


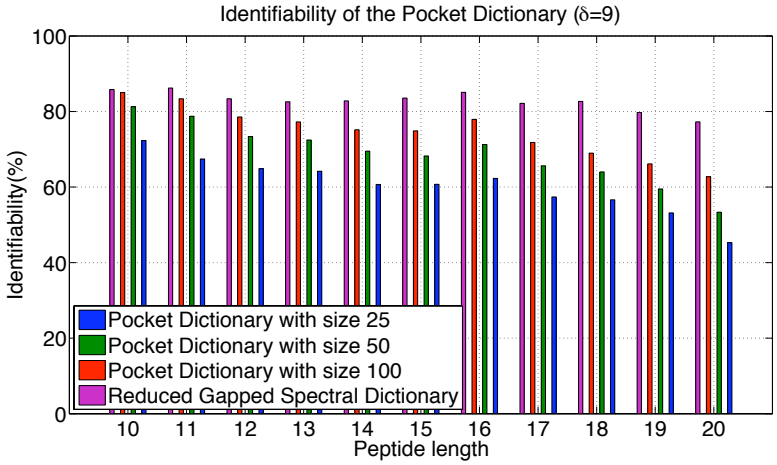
Fig. 8.



(a)



(b)



(c)

Fig. 9.

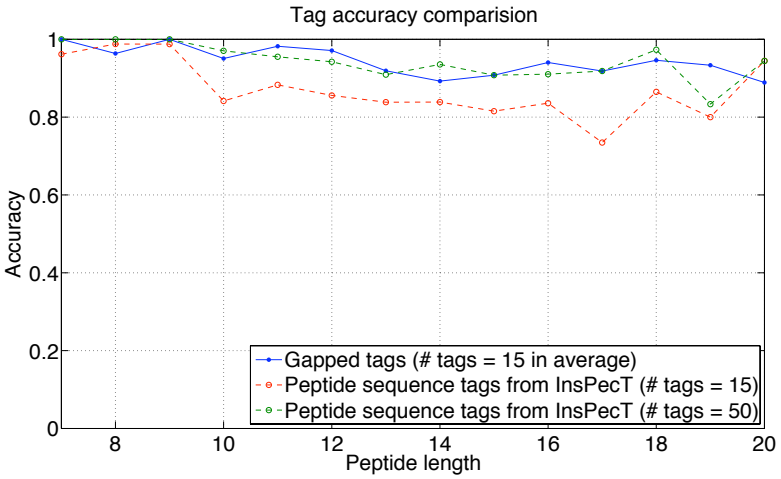


Fig. 10.

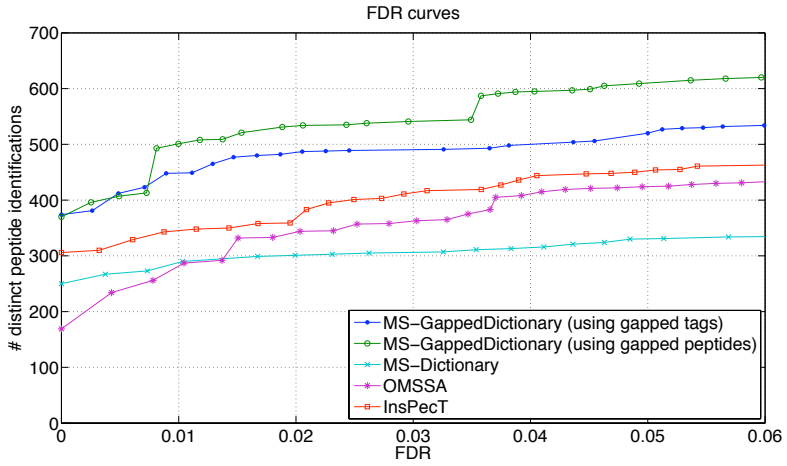


Fig. 11.

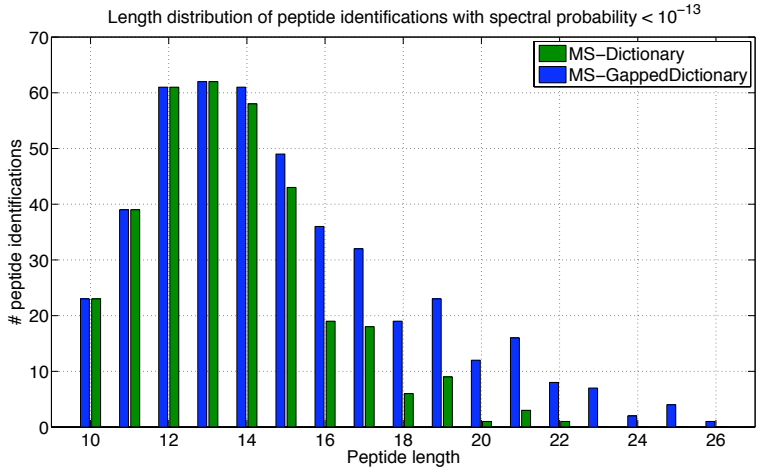


Fig. 12.

No.	Gapped Peptide (GP)	Coverage of GP (%)*	# of peptides represented by GP
1†	[227]RVSQGK	45.69	12
2	[128] [255]VSQGK	15.99	32
3	[128]VRVSQGK	13.71	20
4	[128]VR[186]QGK	11.42	4
5	[128]VRV[215]GK	5.71	2
6†	[383]VSQGK	5.71	2
7	[128]G[198]VSQGK	1.77	20
Total	·	100	92

(a)

No.	Tag	Coverage of tag(%)	Covered GP
1	$\underleftarrow{569}$ QGK $\underrightarrow{0}$	94.3	1,2,3,4,6,7
2	$\underleftarrow{383}$ VSQ $\underrightarrow{185}$	82.9	1,2,3,6,7
3	$\underleftarrow{482}$ SQG $\underrightarrow{128}$	82.9	1,2,3,6,7
4	$\underleftarrow{227}$ RVS $\underrightarrow{313}$	59.4	1, 3
5	$\underleftarrow{128}$ VRV $\underrightarrow{400}$	19.4	3,5

(b)

No.	Gapped Peptide (GP)	Coverage of GP (%)*	# of peptides represented by GP
1	[445][250]S[186]LK	33.81	3286
2†	[695]S[186]LK	19.18	1703
3	[445][337][186]LK	13.28	255
4	[445][250][273]LK	7.67	178
5†	[782]GELK	6.10	684
6†	[695]SGELK	5.55	5563
7	[445][250]S[299]K	4.20	901
8	[445][250]SGELK	3.78	3437
9	[445][337]GELK	1.98	1072
10	[445][250]SG[242]K	1.61	3942
11†	[695]SG[242]K	0.91	1614
12	[445][394]ELK	0.91	507
13	[445][250]SG[370]	0.66	604
14	[445][250][144]ELK	0.20	91
15†	[695][144]ELK	0.07	35
16	[445][337]G[242]K	0.09	162
Total	·	100	24034

(c)

Table 1.