A fast coarse filtering method for protein identification by mass spectrometry

Smriti R. Ramakrishnan [a], Rui Mao [a], Aleksey A. Nakorchevskiy [c], John T. Prince [b], Willard S. Willard [a], Weijia Xu [a], Edward M. Marcotte [b], Daniel P. Miranker [a]

[a]Department of Computer Sciences, University of Texas at Austin, [b]Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology, University of Texas at Austin, [c]Department of Chemistry and Biochemistry, University of Texas at Austin

ABSTRACT

Motivation: We reformulate the problem of comparing mass-spectra by mapping spectra to the vector space model commonly used in document retrieval. It follows that measures of document similarity and document indexing may be adapted for protein identification. In our approach a fast coarse filtering method leveraging a metric space indexing algorithm is used to produce an initial candidate set. We then rank the spectra in this reduced set using ProFound's Bayesian scoring scheme. Ideally,the complexity of the coarse filter search approaches O(log n), as compared to the linear performance provided by most leading tools in the field.

Results: We consider three distance measures based on cosine and hamming distances, modifying them to accommodate the peak shifts intrinsic to mass spectra and investigate their integration with the multivantage-point index structure. Of these, a semi-metric, fuzzy-cosine distance using peptide mass constraints performs the best. We implement an approximate semi-metric search, and show that this improves index pruning power over a standard metric space search.

We measure accuracy of results and index performance on a test set of peptide fragmentation spectra from *E.coli* proteins. We also report sensitivity(recall) and specificity(precision) scores on a more comprehensive benchmark of 1000 Angiotensin-II tandem mass spectra, showing that, in practice, approximate searches in this high dimensional sparse space are acceptable when accompanied by substantial increase in search efficiency.

1 INTRODUCTION

Proteomics experiments are often hindered by the computational expense of protein identification via database lookup of peptide fragmentation spectra. For example, typical analyses of an LC/LC/MS/MS experimental data set using the popular Bioworks program (ThermoFinnegan) on a single processor takes on the order of half a day of computation time. As this is comparable to the time required for data collection, the computation effectively doubles the sample analysis time.

High throughput methods for *in silico* identification of fragmentation spectra (tandem spectra or MS/MS spectra) are becoming increasingly important, due to fast growing protein and gene sequence databases. Most tools today employ linear scans of large databases (using linear filters on query background information when available, to reduce search time). The search hits are approximate and only meaningful when ranked by a probabilistic or statistical significance/relevance score [31, 16, 21]

There are several reasons why even approximate mass spectra searches are computationally expensive. A naive similarity measure is the Shared Peaks Count (SPC)- a count of common m/z values between two spectra. SPC does not account for small peak shifts intrinsic to mass spectra due to measurement and calibration error of the mass spectrometer. Searches must also account for larger peak shifts caused by post-translational peptide modifications and mutations [22]. A common solution is to add modified copies of each spectrum to the database [31].This is called the *virtual database* approach [22]. There are 200+ known protein modifications [13], and this method soon results in exponential blowup of database size due to combinatorial explosion. This method clearly does not scale and linear scans become even more unacceptable. As an alternative, Pevzner et al. [22, 22] proposed an $O(n^2k)$ dynamic programming distance measure that can match two n-dimensional spectra that are up to k peak modifications apart. In the context of current approaches that use linear scans of large databases (size D), this measure must be evaluated for every entry in the database (total time complexity of $O(n^2kD)$).

We propose a fast coarse filtering search algorithm for protein identification that can alleviate many of the problems mentioned above. A coarse filter screens out unlikely candidates at early stages in the search, thus saving unnecessary comparisons. Coarse filtering algorithms have been applied

successfully to genomic databases ([30], [8]). In mass spectra based protein identification, approaches that combine the virtual database approach with complex distance functions similar to Pevzner et al. start to become feasible in the presence of sublinear coarse filtering. Scalable coarse filtering will result in faster searches of larger databases. It may also improve overall quality of search by allowing the use of more discriminative, computationally expensive measures on the reduced candidate set.

We present a 'coarse filtering-fine ranking' scheme for protein identification. Our search methodology consists of a coarse filtering stage that improves on the shared peaks count, followed by a post processing fine-ranking stage. We implement a version of ProFound's [32] Bayesian scoring scheme as an example of a fine filter. The coarse filter overcomes the deficiencies of the shared peaks count while speeding up the search. The fine ranking ranks the results returned by the coarse filter, attaching a probabilistic significance score to each returned result. The system reports the top n matches as computed by the fine filtering rank.

Coarse filtering algorithms for genome databases have traditionally drawn inspiration from text [12] and image retrieval [28]. We describe a fast coarse filtering search method for proteomics based on metric space indexing, leveraging the vector space model from information retrieval. We represent mass spectra as vectors of mass/charge (m/z) values, creating a search space similar to sparse high dimensional document vector spaces. Matching similar images is also often accomplished by comparing high dimensional histograms of image color (frequency spectra). However, due to the discrete nature of both m/z values in mass spectra and word frequency values in document vectors, text retrieval was a better motivation for this system.

We consider three distance measures for comparison of mass spectra. The first is derived from the cosine similarity measure, and adapted to account for peak shifts in experimental spectra. The second, fuzzy cosine distance with peptide precursor mass constraints, achieves maximum reduction in search time. We also investigate hamming distance on reduced dimension boolean spectra vectors. We present an empirical evaluation of the different distance functions, based on retrieval time and accuracy of results. We show that number of distance calculations were reduced to 0.2% of the database size and the candidate set for fine filtering was reduced to 0.11% of the entire database.

Metric space indexing in high dimensional spaces is difficult because nearest neighbor and range query [6] algorithms have an exponential dependency on the dimension of the space [7]. This is known as the *curse of dimensionality* [6]. In our case, a semi-metric distance function is most effective at reducing search time by effectively reducing the intrinsic dimensionality of the space. We find that semi-metric searches on a multiple vantage point (MVP) index tree may be approximate, but achieve better search efficiency (pruning). As the indexing method serves as a coarse filter, and the speedup is substantial, the approximate nature of the search may be acceptable as measured by recall-precision scores.

To summarize, we propose a fast, coarse filtering search for peptide fragmentation spectra. Using semi-metric searches on multiple vantage point trees, we show substantial reduction in search complexity over linear scans, while maintaining quality of results measured using standard precision-recall scores. Our results are ranked by an implementation of ProFound's scoring scheme.

Section 2 gives a brief overview of metric space indexing and protein identification by mass spectrometry. Section 3 details our distance functions for spectra comparison. Section 4 introduces semi-metric searches on MVP trees, and describes evaluation measures for the same. We present experimental results in Section 5 and conclude in Section 6.

2 RELATED WORK

A mass spectrum is a histogram of constituent mass over charge (m/z) ratios of a set of molecules. In bottom-up proteomics, the spectra are derived from peptides generated from the enzymatic digestion of a protein. The m/z value of each peptide is measured by a high precision mass spectrometer. It has been shown that given a sufficient number of accurately measured m/z peaks, a protein can be identified within acceptable statistical significance scores [32]. Closely related to the peptide mass fingerprint (PMF) spectrum, is the peptide fragmentation fingerprint (PFF) spectrum. The induced fragmentation of a single peptide at the peptide bonds, often via collision with inert gas, results in the fragmentation spectrum. Thus fewer, but more precise, fragmentation spectra can uniquely identify the protein. However, especially in MS/MS, automated searches must account for calibration errors, post-translational peptide modifications and mutations which introduce peak shifts into the experimental spectra.

Several approaches to *in silico* identification using MS have been described in the literature. The simplest similarity measure for spectra is the Shared Peaks Count (SPC). A peak is one measured m/z value and the intensity of occurrence. Using SPC alone as a measure of similarity introduces various problems. As already stated, while SPC is an intuitive measure of similarity, its accuracy diminishes quickly in the presence of peak shifts due to mutations and/or modifications [22].

ProFound [32], MASCOT [21] and MS-FIT [9], popular tools for protein identification using peptide mass fingerprinting, use statistical or probabilistic scoring schemes that improve on the shared peaks count. MASCOT and MS-FIT are based on the MOWSE score [20]. MOWSE is a scoring scheme that uses the normalized distribution frequency of peptides in the sequence database. MASCOT reports statistical significance levels and expect values for the MOWSE score ProFound uses a Bayesian scoring scheme. ProFound

gives the largest number of correct identifications as reported in a recent survey of the three systems [2]. Popular tools for MS/MS identification are TurboSEQUEST [31] and MASCOT [21].

Pevzner et al [22] proposed a similarity measure for fragmentation spectra, using a dynamic programming algorithm ($O(n^2k)$), to identify spectra that are at most k modifications/mutations apart. Applying a *band optimization* technique [26] on dynamic programming, could reduce time complexity. Band optimization has been for matching gene sequences [3] and in speech recognition using dynamic time warping (DTW) [25]. We believe our fuzzy cosine distance search space is very similar to the band optimization search space. We are investigating a proof of correctness based on this fact.

2.1 Metric Space Indexing

A non-negative distance function $D_{met}(v_1, v_2)$ that satisfies the following conditions is known as a metric.

1. $D_{met}(v_1, v_2) = 0$ iff$v_1 = v_2$ (identity)
2. $D_{met}(v_1, v_2) = D_{met}(v_2, v_1)$ (symmetry)
3. $D_{met}(v_1, v_2) + D_{met}(v_2, v_3) \geq D_{met}(v_1, v_3)$ (triangle inequality)

A metric space (X, ρ) is defined by a non-empty set X and a metric distance ρ. A distance function that satisfies the identity and symmetry requirements but fails the triangle inequality is called a *semi-metric*. A distance function that satisfies symmetry and the triangle inequality but fails the identity requirement in one direction is called a *pseudometric*. A function with both these properties is called a semi-pseudometric. In this paper, we use semi-pseudometric interchangeably with semi-metric.

Objects in (non-linear) metric spaces need not have a geometrical representation, i.e, there need not be a zero point or origin. In n-dimensional real vector spaces an object is a point in $\Re^n$ space and has a geometric meaning. If mass spectra are represented as lists of m/z values, they form a $\Re^n$ vector space. Vector spaces like $\Re^n$ or even Boolean space $(0, 1)^n$, along with distance metrics like the L_p norm (Minkowski distance $[\sum_i |a_i - b_i|^p]^{\frac{1}{p}}$) or the cosine similarity ($\frac{\sum_i a_i.b_i}{\sqrt{\sum_i a_i^2}\sqrt{\sum_i b_i^2}}$) are a subset of metric spaces.

A range query on a metric space will return all points u of a given distance r from a query point q, such that D(u, q) $\leq$ r. By leveraging the triangle inequality, an index built over a metric space avoids distance computations with points that are unlikely to be within radius r of the query. Metric space indexing thus reduces search time by decreasing the number of runtime distance computations. In a pivot based index structure [6], the search space is partitioned into disjoint regions recursively. In each recursion, one or more pivots (vantage-point or VP) are first selected. Then, the data points are partitioned into two (or more) disjoint branches based on their distances from the pivot(s). MVP-Trees [1] extend VP-Trees by increasing the number of disjoint datasets into which a dataset is partitioned.

3 DISTANCE MEASURES FOR COMPARISON OF MASS SPECTRA

In this section, we introduce three distance functions for mass spectra. We draw inspiration from the vector space model in text retrieval and the shared peaks count in mass spectrometry. Documents are commonly represented as sparse, high dimensional vectors, where the i^{th} entry represents a measure of occurrence-frequency of the i^{th} word. We need distance measures that will improve on the shared peaks count (by detecting small and large peak shifts), and also act as good coarse filters.

We first introduce a vector space data representation for spectra, investigating both coarse resolution and high resolution vectors. We then define three distance measures that are theoretically able to account for peak shifts due to both calibration error and mutation/modification. We investigate metric properties of each distance function in Section 3.4.

3.1 Data representation

A peptide fragmentation spectrum is a histogram of mass over charge (m/z) ratios versus intensity. It is common practice to use only m/z peak lists, ignoring intensity information [22]. Given a m/z range of $[M_1, M_2]$ Dalton (Da) and resolution of representation M_{res} Da, mass spectra can then be visualized as sparse boolean vectors, where a non zero entry signifies a peak at that m/z value (or if the resolution of representation is $M_{res} > 1.0$, the presence of a peak in the range of m/z values represented by the i^{th} dimension). Visualizing each spectrum as a boolean vector allows us to reformulate the problem as a metric space indexing problem. It must be emphasized that though this explanation deals with equi-sized boolean vectors - the actual implementation deals with *non-boolean* compressed vectors using m/z values directly. A fixed length boolean vector analogy is useful here as it allows us to derive a distance function using principles from document retrieval.

A coarse filter serves two purposes: to reduce the number of distance computations and be discriminative enough to return a small relevant resultset. Any combination of data representation and distance metric must intuitively ensure that we count peaks that differ by known amounts. Peak shifts due to calibration error are small, in the range of 0-1Da, whereas common modifications can cause large peak shifts from 50-200 Da. We hypothesize that this peak shift can be handled either by the data representation or by the distance metric. Section 3.2 describes a high resolution (high dimension) data representation with *fuzzy* cosine distances. Section 3.5 describes a coarse resolution (low dimension) data representation with an exact Hamming Distance distance metric.

3.2 Fuzzy Cosine Distance

We show that the cosine similarity measure from text retrieval can be rewritten as a length-normalized Shared Peaks Count. Then we define a *fuzzy* cosine measure, and show that by varying a peak mass tolerance factor, τ_{ms}, we can account for peak shifts.

Given a mass spectrum P (a list of m/z value peaks) and resolution $0 < M_{res} \leq 1.0$ Da, define a high dimensional boolean vector S such that

$$S[i] = \begin{cases} 1 & \exists p \in P, (p - M_{res} * i) \leq M_{res}, and \\ & (p - M_{res} * (i-1)) > M_{res} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

The second condition ensures that each peak in a spectrum maps to only one non-zero entry in S. The spectrum can be visualized as a boolean vector in $(0,1)^N$ space, N = $(M2 - M1 + 1)/M_{res}$. For example, given a m/z range of 100-5000 Da and $M_{res} = 0.1$ Da, we are looking at sparse, 49,000 dimension vectors.

Given a peak mass tolerance, τ_{ms}, such that $\tau_{ms} \geq M_{res}$, we can define range $k = \tau_{ms}/M_{res}$. Shared Peaks Count within a tolerance window, using range k, can then be defined as

$$SPC_\tau(A,B) = \sum_i match(a_i, b_j); j \in [i-k, i+k] \quad (2)$$

$$match(a_i, b_j) = \begin{cases} 1 & a_i = b_j = 1 \\ & match(a_m, b_j) = 0, m \in [1, i) \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Again the second condition ensures that one peak can only count for one match - multiple matches are not allowed. We observe that for zero peak tolerance, $\tau_{ms} = 0$, the shared peaks count reduces to the dot product on boolean vectors.

$$SPC_\tau(A,B) = \sum_i match(a_i, b_i) = A.B \quad (4)$$

We also note that cosine similarity is defined as the normalized dot-product between two vectors.

$$Cos(A,B) = \frac{A.B}{\|A\|\|B\|} \quad (5)$$

where $\|A\|$ is the L2 norm over vector A. Modifying Equation 5 for $\tau_{ms} > 0$, we define a *fuzzy* cosine similarity measure

$$Cos_\tau(A,B) = \frac{SPC_\tau(A,B)}{\|A\|\|B\|} \quad (6)$$

Finally, since a metric space index requires a *distance* metric, we define fuzzy cosine distance as the inverse cosine of Cos_τ.

$$D_{ms}(A,B) = \arccos(Cos_\tau(A,B)) \quad (7)$$

Given our boolean representation of protein mass spectra, and since a metric space is a generalization of a vector space [6], cosine distance is one obviously applicable distance function for metric space indexing of mass spectra. Another compelling reason is the observation that the numerator of cosine distance for boolean vectors is the same as the shared peaks count.

3.3 Tandem Cosine Distance

Tandem cosine distance combines fuzzy cosine distance with the precursor mass of the query peptide. Peptides with vastly differing precursor mass are unlikely to be similar, are should be further apart in vector space. We factor a corresponding precursor mass difference term into the fuzzy cosine distance. Given two peptide sequences A, B and precursor masses M_A, M_B; we define tandem cosine distance D_{tcd} as

$$D_{tcd}(A,B) = D_{ms}(A,B) + D_{pm}(A,B) \quad (8)$$

D_{pm} is a distance function that computes absolute difference in precursor mass within a tolerance window. In order to account for slight differences in analytical and experimentally measured precursor mass, we introduce a precursor mass tolerance factor, τ_{pm} and define D_{pm} as

$$D_{pm}(A,B) = \begin{cases} 0 & |M_A - M_B| \leq \tau_{pm} \\ |M_A - M_B| & \text{otherwise} \end{cases} \quad (9)$$

There are two reasons why it is important to include precursor mass into the distance function. First, current MS/MS search tools filter the database on precursor mass *first* using a *linear* scan and then apply more expensive distance measures. Our goal is a sublinear coarse filter that combines precursor mass and peak list similarity into one distance function. Second, tandem cosine distance reduces search time drastically when compared to simple fuzzy cosine distance - it is a semi-metric and a better coarse filter for reasons detailed in Section 4.

3.4 Metric properties of modified cosine distances

Fuzzy cosine distance is a semi-pseudometric distance function. As a consequence of the tolerance window, fuzzy cosine distance may not satisfy the triangle inequality and it may not always satisfy the identity criterion in both directions (proof omitted). By the additive property of metric spaces, tandem cosine distance, D_{tcd} is also a semi-pseudometric. Similarly, it can be shown that D_{pm} is also a semi-pseudometric.

3.5 Hamming Distance

Hamming Distance is defined as the cardinality of XOR(V_1, V_2). Intuitively it counts the number of mismatched peaks, and is a distance metric. We use an overlapping window to generate coarse resolution boolean vectors. If $M_{res} > 1.0$ is the window size, and S is the window overlap size, size of vector V is N = $(M_1 - M_2)/S$. Here V[i] = 1 iff $\exists$ peak p, p $\in [M_1 + (i-1)S, M_1 + iM_{res}]$.

Though coarser resolutions reduce the number of distance computations (indexing of low dimensional vectors is an

easier problem), the number of results returned increases drastically with increase in window size. This is because the probability of a random match increases with coarser resolutions. Since tandem cosine distance gave us better search efficiency, this paper does not elaborate further on coarse resolution hamming distance methods. However, the approach is promising, especially when we observe that using coarser resolutions (100-200 Da) is a simple way of matching larger peak shifts for detecting mutations/modifications.

4 SEMI-METRIC SEARCH

4.1 Reducing the intrinsic dimensionality

The dimensionality of a space is not easily defined, especially for metric spaces. An alternative is to define the *intrinsic dimensionality* [6] as $\rho = \frac{\mu^2}{2\sigma^2}$ where μ and σ^2 are the mean and variance of the histogram of pairwise distances between points in the space. In other words, in a plot of pairwise distances, a large mean and/or low variance implies a high intrinsic dimensionality.

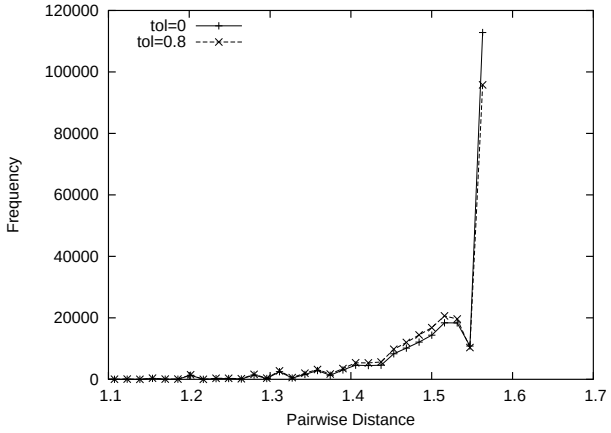

Fig. 1. Pairwise distance distribution of database (exact and fuzzy cosine distance)

Algorithmic performance degrades exponentially with increase in intrinsic dimensionality. This has been referred to as the *curse of dimensionality*. A good discussion can be found in Chavez and Navarro [6]. Due to the high intrinsic dimensionality of the search space, an exact metric space solution to our problem suffers from the curse of dimensionality and is only slightly more efficient than a linear scan. This phenomenon has also been observed in document vector spaces [27]. Pairwise distance histograms of exact cosine distance (Figure 1) show a large mean and variance on mass spectra space, corresponding plots (Figure 1, Figure 2)for fuzzy and tandem cosine distances (semi-metrics) show lower means. A semi-metric distance function actually has the effect of *reducing* the intrinsic dimensionality of the search space. Having

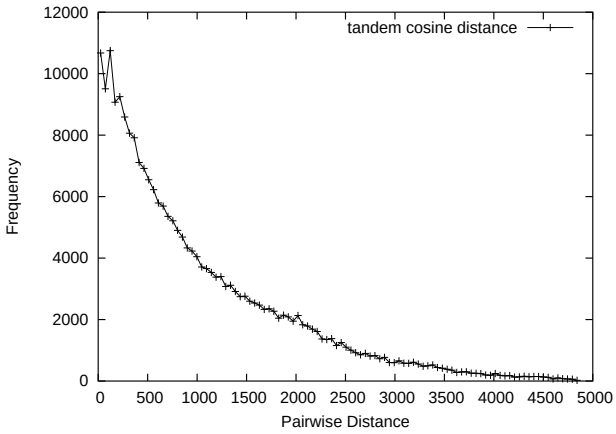

Fig. 2. Pairwise distance distribution of database (tandem cosine distance)

justified that a semi-metric search might be more suitable in this coarse filtering application, it follows that a semi-metric function must be integrated with the metric space index.

4.2 Modifying the index for a semi-metric search

Given a pivot p_i and a query (q, r) in a metric space search of radius r, we would prune all points u such that

$$|d(u, p_i) - d(q, p_i)| > r \tag{10}$$

A semi metric distance function fails the triangle inequality, by some amount k ($d(X, Y)+d(Y, Z)+k \geq d(X, Z)$). In this case, Chavez and Navarro [5] show that, there may exist some u such that d(q,u) + k >r, but d(q,u)<r. This means some points u may be incorrectly pruned. However, if we can bound the amount by which the triangle inequality will fail, the metric space index equations can be *adjusted* to return exact results. We briefly describe this procedure for the case of (multiple) vantage point (MVP) index trees. In this case, for an exact semi-metric search, Equation 10 is modified to

$$|d(u, p_i) - d(q, p_i)| > (r + \kappa) \tag{11}$$

For fuzzy cosine distance, we can derive (proof omitted) a very loose upper bound on κ, when *every* peak in one vector differs from its corresponding match in the other vector by the peak tolerance τ_{ms}.

$$\arccos(\frac{N-1}{N}) + 2\tau_{pm} \leq TOL \leq \frac{\pi}{2} + 2\tau_{pm} \tag{12}$$

Our hypothesis is that using κ in practice might be overkill, and is likely to result in a large number of false positives due to conservative pruning. Using $TOL \leq \kappa$ is a more aggressive pruning technique. In practice it is difficult to determine TOL. By using $TOL = \tau_{ms} + \tau_{pm} \leq \kappa$ we theoretically revert to an approximate search. However, our results show that by choosing a suitable value of TOL, we can keep the precision of results at 90%. Using $TOL \leq \kappa$ gives near 90% precision, while achieving 99% pruning of the database. It will be useful

Table 1. Databases and test sets

Test	Database Size	Test set Size	Acceptable Radius
Search Efficiency	137,349	14 (*E. coli*)	1.82
Search Quality	138,341	992 (Angiotensin-II)	1.56
Scalability	653,882	14 (*E. coli*)	1.82

Database and test set size in terms of number of spectra.
Table 2. Databases and test sets

Table 3. Parameters for theoretical digest of *E. coli* proteins

Number of Proteins	4824
m/z Peak Tolerance	0.2 Da
Precursor Mass Tolerance	2.0 Da
Charge State	+1
Ion States	b, y
Missed Cleavages	0
Enzyme	Trypsin
Mass Range	0-5000 Da

to derive probabilistic bounds on the correctness of the search in the future.

4.3 Evaluation of Semi-Metric Searches

Recall(sensitivity) and Precision(specificity) are frequently used to measure the quality of approximate searches. Ideally, we want to maximize both Precision and Recall. Here TP stands for number of true positives, FN is the number of false negatives and FP is the number of false positives.

$$R_{tcd} = Recall(Sensitivity) = TP/(TP + FN) \quad (13)$$

$$P_{tcd} = Precision(Specificity) = TP/(TP + FP) \quad (14)$$

5 RESULTS

This section describes our experimental methodology and results. We ran range and k-nearest neighbor queries on a multiple vantage point tree index structure modified to incorporate semi-metric searches. Section 5.1 reports Search Efficiency (Number of distance computations and size of candidate set at acceptable radii). Section 5.2 measures the quality of semi-metric search, reported using Recall-Precision pairs. Section 5.3 reports scalability results.

The database and test sets used for each test are summarized in Table 2, along with acceptable radii for that test. It is nearly impossible to acquire unambiguously identified spectra from a complex sample, as the only current means for verifying a protein's identity from a mass spec analysis lies in software whose accuracy is still under question. We define 'correctness' by comparing our top hit (after fine filtering) with the top hit from the TurboSEQUEST [31]. This high confidence result is expected to be correct because it also had high protein and peptide probability scores after analysis with ProteinProphet and PeptideProphet [18, 14]. There seems to be good justification for trusting the ProteinProphet and PeptideProphet probabilities, especially for high probability identifications.

In all queries, our fine filter ranked the correct answer as the top hit, with an identification probability of > 99% in most cases. The scores between first and second ranked peptides differed by at least three orders of magnitude (up to eight orders of magnitude in many cases). We report results at the radius at which all 'correct' identifications were returned by the coarse filter.

5.1 Index performance

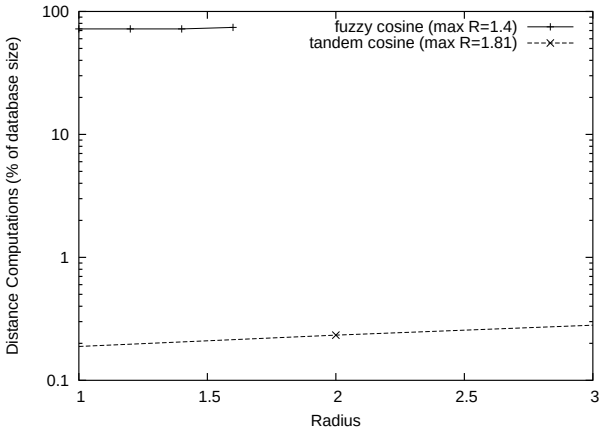

Fig. 3. Tandem Cosine Distance Vs Fuzzy Cosine Distance: Percent of database searched at acceptable radii R

This section compares fuzzy cosine distance, tandem cosine distance and hamming distance in terms of the average number of distance computations and size of candidate set. The test database consisted of the 4894 proteins derived from the genome of Escherichia coli K12 (*E. coli*), a subset of the SWISSPROT database from UNIPROT version 45.0. These proteins were theoretically digested into 137,349 predicted spectra. The test set consisted of 14 experimental tandem mass spectra chosen from the Open Proteomics Database [23], accession number opd00006_ECOLI. The digest parameters are given in Table 3.

Tandem cosine distance performs the best - both in terms of percentage of database that is searched and the size of the filtered resultset. Figure 3 shows the percentage of database searched for both cosine distance based measures. Figures 4 shows a linear increase in the number of distance computations and the size of the candidate set for tandem cosine

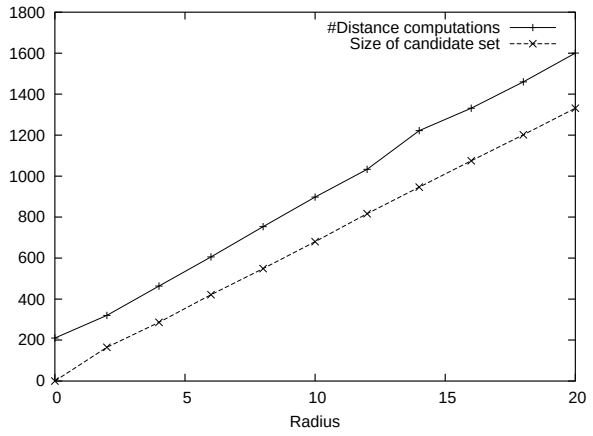

Fig. 4. Number of Distance Computations and Size of Candidate Set Vs Radius (Tandem Cosine Distance, Database=137,349 spectra)

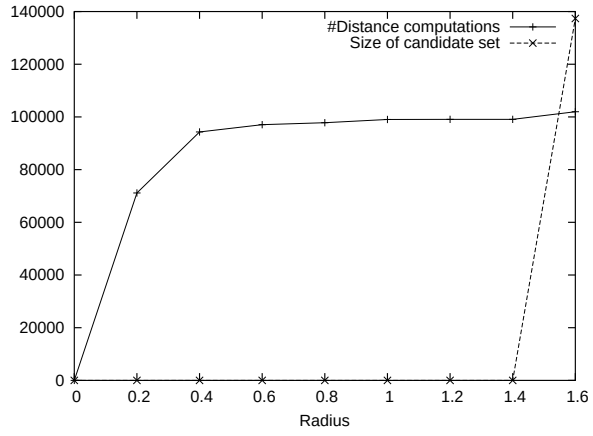

Fig. 5. Number of Distance computations and Size of Candidate Set Vs Radius (Fuzzy Cosine Distance, Database=137,349 spectra)

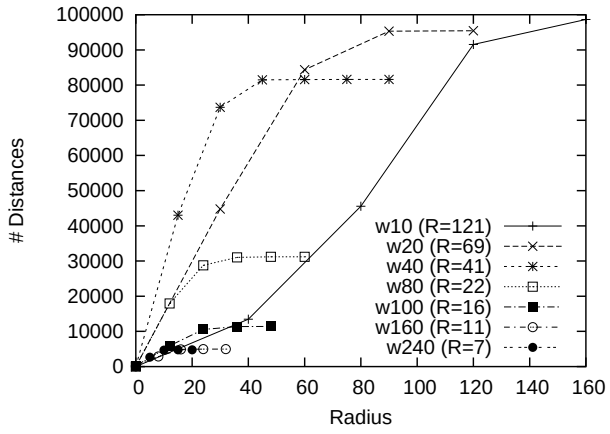

Fig. 6. Number of distance computations Vs Radius (Hamming Distance, Database=137,349 spectra)

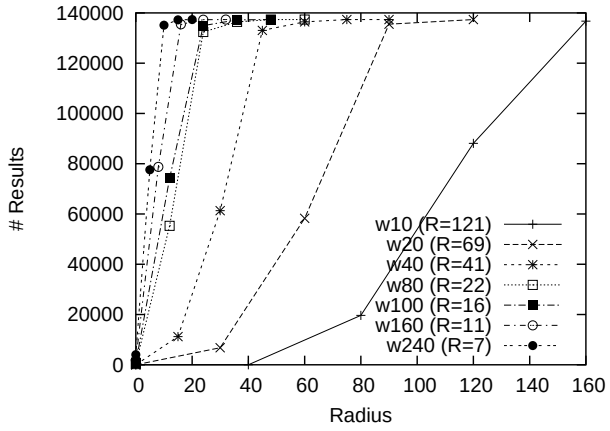

Fig. 7. Number of results returned Vs Radius (Hamming Distance, Database=137,349 spectra)

distance. Tandem cosine distance searches about 0.22% of the database and returns only about 0.1% of the database. Fuzzy cosine distance (Figure 5) performs worse, returning less than 0.003% of the database after searching almost 75% of the database. Similar numbers for Hamming distance are shown in Figures 6 and 7. As mentioned in Section 3.5, the number of results returned by a hamming distance index is very large as compared to the reduction in distance computations. Though the percentage of the database searched reduces with increase in window size, correct results are returned at larger radii, and the size of the candidate set remains large.

5.2 Accuracy

This section describes quality measurements for the approximate semi-metric search. We report recall-precision scores for the Angiotensin-II benchmark using tandem cosine distance. Experimental Angiotensin-II fragmentation spectra were collected on the LCQ Deca XP Plus ion trap mass spectrometer running Xcalibur data acquisition software. A total of 1000 MS2 scans were collected. The complete experiment details are available online [29]. The test database was created by adding 992 Angiotensin-II experimental spectra into the *E. coli* database described in Section 5.1..

Angiotensin-II is a set of 992 experimentally generated spectra from the same peptide. Each spectrum in the benchmark is different, but similar enough to be recognized as the same peptide. We plotted a histogram of pairwise distances on the query set and computed the average (R=1.42) and maximum (R=1.56) query radii from this plot. We measure the ability of the distance measure to return all 992 (recall) and only 992 (precision) spectra per query.

Figure 8 is a recall-precision plot for different distance measures. Choosing a 'good' value for TOL (parameter by which index equations are adjusted) makes the recall-precision plot near ideal. Recall, precision, number of distances and number

Table 4. Angiotensin-II benchmark: Tandem Cosine Distance

Radius	Recall	Precision	#Distances	#Results
0.0	0.001	1.0	1345.0	1.004
0.5	0.001	1.0	1474.0	1.004
1.0	0.001	1.0	1475.0	1.034
1.42	0.5023	0.9962	1582.82	499.44
1.5	0.8537	0.9644	1587.0	878.39
1.56	0.9373	0.8751	1587.0	1067.26
1.6	1.0	0.7936	1587.0	1250.0
2.0	1.0	0.7708	1587.0	1286.98
3.0	1.0	0.7078	1691.0	1401.57
5.0	1.0	0.6124	1913.0	1619.85

Acceptable Radius R = 1.56

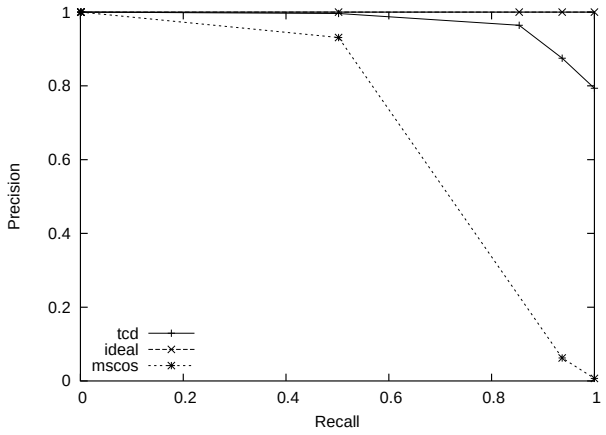

Fig. 8. Recall-Precision curves for the Angiotensin-II benchmark

of results for different radii are shown in Table4. At the maximum radius (R=1.56) for tandem cosine distance, precision approaches 90% and 95%recall. This validates our assumption that aggressive pruning, at the cost of a theoretically approximate search, might yield good benefits in terms of search efficiency.

5.3 Scalability of the coarse filter

The database consisted of 653,882 predicted spectra from 4279 *E. coli* proteins and 19821 Human proteins (datasets available online at [19]). The test set is the same as used in Section 5.1. A series of different size databases was built from the dataset. For each database, a set of k-NN queries was executed with k=100. Moreover, the query results are bounded by a radius(R) to the query object. The MVP tree implementation is part of MoBIoS [11], a special purpose database management system for molecular biology. Although in this study the system is main-memory based, the MVP-tree is organized for pagination to disk. Like the depth of a B+ tree in a relational database system, the MoBIoS MVP tree has discontinuous

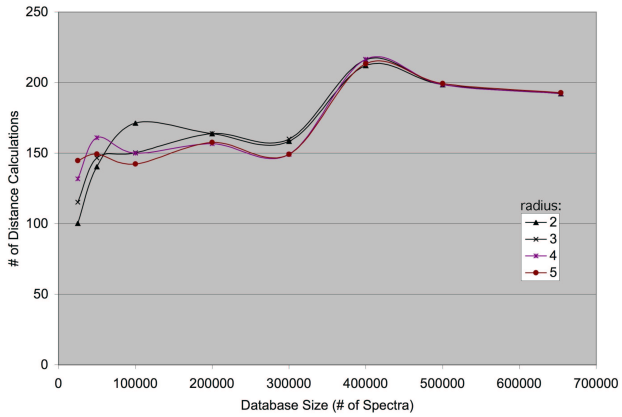

Fig. 9. Scalability: k-NN queries (k=100) showing Number of distance computations varying with Database Size

increases in height as the database grows. Thus, the performance degrades very slowly, subject to sudden increments when the index increases height (for example from 300,000 to 400,000). Figure 9 shows only a slight increase in the average number of distance calculations as database size increases. Since the single correct answer for our application is determined by the fine ranking phase, the topmost hit may not be closest to the query, especially as database size grows. This behavior can be countered by empirically choosing an appropriate value for k.

6 DISCUSSION AND FUTURE WORK

We described a fast coarse filtering-fine ranking scheme for peptide fragmentation spectra using metric space indexing. We showed that a semi-metric fuzzy cosine distance with precursor mass constraints achieves maximal reduction in both the number of distance comparisons (0.2% of database) and the size of the candidate set (0.11% of database). At acceptable radii, we reported 90% average precision on a 1000 protein Angiotensin-II benchmark. We also showed scalability of the coarse filter on k-NN queries on the modified MVP index. A basic version of the system for protein identification via peptide mass fingerprinting is accessible online [17].

We offered some solutions to the automatic detection of mutations and modifications. The exponential blowup, caused by adding extra theoretically modified spectra into the database, will have less drastic effects in search time due to a sublinear coarse filter. Alternatives to the virtual database approach are also made more feasible. One possible solution to matching mutations/modifications is to derive a distance metric from Pevzner's dynamic programming similarity measure algorithm and use it as the distance measure for a coarse filter- the $O(n^2k)$ complexity will be countered by sublinear search complexity. As mentioned in Section 2, there exists room for decreasing the time complexity of Pevzner

et al.'s dynamic programming approach using band optimization techniques [25]. Another alternative, is to use coarse resolution hamming distance with precursor mass constraints to detect mutations/modifications. This would be similar to increasing the tolerance window of fuzzy cosine distance. We believe that either of these coarse filters will return a candidate set that is a super set of Pevzner et al.'s resultset and we are working on deriving a theoretical proof of this behavior.

We plan to test our system against identified semi-complex spectra, few of which are publicly available [15, 24, 10]. We would also like to investigate the direct integration of a probabilistic ranking scheme with the results returned by the index [4].

REFERENCES

[1]T Bozkaya and M Ozsoyoglu. Distance-based indexing for high-dimensional metric spaces. In *SIGMOD '97: Proceedings of the 1997 ACM SIGMOD international conference on Management of data*, pages 357–368. ACM Press, 1997.

[2]DC Chamrad, G Korting, K Stuhler, HE Meyer, J Klose, and M Bluggel. Evaluation of algorithms for protein identification from sequence databases using mass spectrometry data. *Proteomics*, 4(3):619–628, 2004.

[3]KM Chao, WR Pearson, and W Miller. Aligning two sequences within a specified diagonal band. *Comput. Appl. Biosci.*, 8(5):481–487, 1992.

[4]S Chaudhuri, G Das, V Hristidis, and G Weikum. Probabilistic ranking of database query results. In *Proceedings of the 30th VLDB Conference*, pages 888–899, 2004.

[5]E Chavez and G Navarro. A probabilistic spell for the curse of dimensionality. In *ALENEX: International Workshop on Algorithm Engineering and Experimentation, LNCS*, 2001.

[6]E Chavez, G Navarro, R Baeza-Yates, and JL Marroquin. Searching in metric spaces. *ACM Comp. Surv.*, 33(3):273–321, 2001.

[7]B Chazelle. Computational geometry: a retrospective. In *STOC '94: Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, pages 75–94. ACM Press, 1994.

[8]W Chen and K Aberer. Efficient querying on genomic databases by using metric space indexing techniques. In *In Proc. of 8th Int. Work on Database and Expert System Applications*. IEEE Computer Society Press, September 1997.

[9]KR Clauser, P Baker, and AL Burlingame. Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing ms or ms/ms and database searching. *Anal Chem.*, 71(14):2871–2882, 1999.

[10]http://sashimi.sourceforge.net/repository.html.

[11]W Xu D Miranker and R Mao. Mobios: a metric-space dbms to support biological discovery. In *Proc. Of the Int. Conf. On Scientific and Statistical Database Management Systems (SSDBM)*, page 241, 2003.

[12]C Faloutsos and D Oard. A survey of information retrieval and filtering methods. Technical report, University of Maryland, College Park, MD, 1996.

[13]A Gooley and N Packer. *Proteome Research: New Frontiers in Functional Genomics*, chapter The importance of co- and post-translational modifications in proteome projects, pages 65–91. Springer-Verlag, 1997.

[14]A Keller, AI Nesvizhskii, E Kolker, and R Aebersold. Empirical statistical model to estimate the accuracy of peptide identifications made by ms/ms and database search. *Anal Chem*, 74(20):5383–5392, Oct 2002.

[15]A Keller, S Purvine, AI Nesvizhskii, S Stolyar, DR Goodlett, and E Kolker. Experimental protein mixture for validating tandem mass spectral analysis. *OMICS*, 6(2):207–212, 2002.

[16]M Mann and M Wilm. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem*, 66(24):4390–4399, Dec 1994.

[17]http://aug.csres.utexas.edu:8080/msfound/index.html.

[18]AI Nesvizhskii, A Keller, E Kolker, and R Aebersold. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.*, 75(17):4646–4658, 2003.

[19]Open proteomics database. http://bioinformatics.icmb.utexas.edu/OPD/.

[20]DJC Pappin, P Hojrup, and AJ Bleasby. Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.*, 3(6):327–332, 1993.

[21]DN Perkins, DJ Pappin, DM Creasy, and JS Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, 1999.

[22]PA Pevzner, Z Mulyukov, V Dancik, and CL Tang. Efficiency of database search for identification of mutated and modified proteins via mass spectrometry. *Gen. Res.*, 11(2):290–299, 2001.

[23]JT Prince, MW Carlson, R Wang, P Lu, and EM Marcotte. The need for a public proteomics repository. *Nature*, 22(4):471–472, 2004.

[24]S Purvine, AF Picone, and E Kolker. Standard mixtures for proteome studies. *OMICS*, 8(1):79–92, 2004.

[25]H Sakoe and S Chiba. A dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1), 1978.

[26]D Sankoff and JB Kruskal. *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison*. Addison-Wesley, 1983.

[27]T Skopal, P Moravec, J Pokorný, and V Snásel. Metric indexing for the vector model in text retrieval. In *SPIRE*, pages 183–195, 2004.

[28]JR Smith and SF Chang. Tools and techniques for color image retrieval. In *Proc. IS&T/SPIE Storage and Retrieval for Still Image and Video Databases IV*, pages 426–437, 1996.

[29]Test data: Angiotensin-ii benchmark sample preparation detail. http://mobios.csres.utexas.edu/msfound/ismb/data.

[30]HE Williams. Cafe: an indexed approach to searching genomic databases. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, page 389. ACM Press, 1998.

[31]JR Yates III, J Eng, AL McCormack, and D Schieltz. Method to correlate tandem mass spectral data of modified peptides to amino acid sequences in the protein database. *Anal. Chem.*, 67(8):1426–1436, 1995.

[32]W Zhang and BT Chait. Profound - an expert system for protein identification using mass spectrometric peptide mapping information. *Anal. Chem.*, 72(11):2482–2489, 2000.