An application of latent topic document analysis to large-scale proteomics databases

Sebastian Klie[1*], Lennart Martens[2*], Juan Antonio Vizcaino[2], Richard Cote[2], Phil Jones[2], Rolf Apweiler[2], Alexander Hinneburg[1], Henning Hermjakob[2]

[1] Institute of Computer Science
Martin-Luther-University Halle-Wittenberg, Germany
{klie,hinneburg}@informatik.uni-halle.de
[2] EMBL Outstation, European Bioinformatics Institute,
Wellcome Trust Genome Campus,
Hinxton, Cambridge, UK
{lennart.martens,juan,rcote,pjones,apweiler,hhe}@ebi.ac.uk

Abstract. Since the advent of public data repositories for proteomics data, readily accessible results from high-throughput experiments have been accumulating steadily. Several large-scale projects in particular have contributed substantially to the amount of identifications available to the community. Despite the considerable body of information amassed, very few successful analysis have been performed and published on this data, levelling off the ultimate value of these projects far below their potential. In order to illustrate that these repositories should be considered sources of detailed knowledge instead of data graveyards, we here present a novel way of analyzing the information contained in proteomics experiments with a 'latent semantic analysis'. We apply this information retrieval approach to the peptide identification data contributed by the Plasma Proteome Project. Interestingly, this analysis is able to overcome the fundamental difficulties of analyzing such divergent and heterogeneous data emerging from large scale proteomics studies employing a vast spectrum of different sample treatment and mass-spectrometry technologies. Moreover, it yields several concrete recommendations for optimizing proteomics project planning as well as the choice of technologies used in the experiments. It is clear from these results that the analysis of large bodies of publicly available proteomics data holds great promise and is currently underexploited.

1 Introduction

The field of proteomics has undergone several dramatic changes over the past few years. Advances in instrumentation and separation technologies [1, 6] have enabled the advent of high-throughput analysis methods that generate large amounts of proteomics identifications per experiment. Many of these datasets were initially only published as supplementary information in PDF format and, while available, were not readily accessible to the community. Obviously, this

situation led to large-scale data loss and was perceived as a major problem in the field [10, 20].

Several public proteomics data repositories, including the Global Proteome Machine (GPM) [2], the Proteomics Identifications Database (PRIDE) [12, 16] and PeptideAtlas [5] were constructed to turn the available data into accessible data, thereby reversing the trend of increasing data loss.

As a case in point, several large-scale proteomics projects that have recently been undertaken by the Human Proteome Organization (HUPO), including the Plasma Proteome Project (PPP) [19] and the Brain Proteome Project (BPP) [9], have published all of their assembled data in one or more of these repositories. As a result, their findings are readily accessible to interested researchers. It is therefore remarkable to see that very little additional information has so far been extracted from the available data. One of the rare examples where the analysis of large proteomics datasets resulted in a practical application is the recent effort by Mallick and co-workers in which several properties of a large amount of identified peptides were used to fine-tune an algorithm that can predict proteotypic peptides from sequence databases [15].

We here present a novel way to reveal the information that lies hidden in large bodies of proteomics data, by analyzing them for latent semantic patterns. The analysis performed here focused on the HUPO PPP data as available in the PRIDE database. Briefly, the HUPO PPP sent out a variety of plasma and serum samples, collected from different ethnic groups and at different locales worldwide. All of the five resulting plasma samples were additionally treated with one of three distinct methods of anticoagulation: EDTA, citrate or heparin. The total amount of distinct samples thus amounts to twenty: five serum samples, and three times five plasma samples [19].

We used the original peptide sequences to evaluate experiment similarity by performing a latent semantic analysis, a technique often employed in natural language processing. Our results suggest that LSA can be considered a useful analysis tool of such data yielding results which cannot easily be obtained by conventional means.

2 Latent semantic analysis

In order to assess inter-experiment similarity in an all-against-all comparison, an information retrieval method called latent semantic analysis (LSA; also referred to as latent semantic indexing, LSI) is employed. The fundamentals of LSA are well understood and it has been widely used for various information retrieval tasks. The general idea of LSA is to map document into some latent semantic space, in which the dimensions consists of latent topics. The main task is, to reduce the documents from a word-based representation to a topic-based representation, which reduces the influence of noise (random words) during the similarity computation between pairs of documents. Applied to the context of proteomics, experiments take the role of documents while peptides identified in one experiment (more precisely their amino acid sequence) act as terms. The

algorithm reports a similarity score for each pair of experiments, based on the latent topics in peptide representation of the experiments.

2.1 Vector Space representation of proteomics data

LSA computes latent topics from a vector representation of the documents (vector space model). A term-document matrix $W \in \mathbb{R}^{n,m}$ represents a document collection of m documents over a vocabulary of n terms. A value $w_{i,j}$ is the number of occurrences of a particular term i (rows of W) within the jth document (columns of W). In case of proteomics experiments the words are peptides detected by mass spectrometry. As no quantitative information about the peptides is available but just the information about the occurrence, the document term matrix is a bit-matrix in this case.

Each column-vector $\boldsymbol{w}_{\cdot,j}$ in W can be interpreted as a document-vector $\boldsymbol{d}_j$ which characterizes a document (proteomics experiment) by its terms (peptides) [22]. The similarity between two documents $\boldsymbol{a}$ and $\boldsymbol{b}$ represented by their documents vectors is determined by cosine similarity, which gives the cosine of the angle between $\boldsymbol{a}$ and $\boldsymbol{b}$:

$$sim(\boldsymbol{a}, \boldsymbol{b}) = \cos\alpha(\boldsymbol{a}, \boldsymbol{b}) = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{\|\boldsymbol{a}\| \cdot \|\boldsymbol{b}\|} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}} \tag{1}$$

The similarity is between zero and one, due to the normalization. To counteract poorly differentiating, often-occurring terms each row of W is weighted by the inverse document frequency (IDF). IDF of a term t is defined as

$$IDF_t = \log \frac{N_D}{d_t} \tag{2}$$

The effect of IDF-weighting is that poorly differentiating, often-occurring terms will have a much lower weight than highly specific, rare terms.

The vector space model has several drawbacks. First, it relies on exact term matching (because of the scalar product in (1)), thus making it impossible for the model to detect synonyms. A proteomics example of a synonym is the substitution of the isobaric amino acids isoleucine and leucine within a peptide sequence. A second problem is the sparseness of non-zero values in the obtained document/term matrix [13]. The matrix is thus largely composed of zeroes, highlighting the fact that many peptides failed to be identified in more than one experiment, which is a common situation in shotgun proteomics experiments.

2.2 Singular Value Decomposition

In order to reduce the sparseness of the document/term matrix and to detect hidden (latent) term relations, LSA projects the original documents vectors into a lower dimensional semantic space where documents which contain repeatedly co-occurring terms will have a similar vector representation. This effectively

overcomes the fundamental deficiencies of the exact term-matching employed in a VSM [14]. As such, LSA might predict that a given term should be associated with a document, even though no such association was observed in the original matrix [4]. The core principle for achieving this is the application of singular value decomposition (SVD), a type of factor analysis which can be applied to any rectangular matrix in the form of:

$$W = U\Sigma V^T \tag{3}$$

where $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{m\times m}$ are orthogonal matrices (i.e. $UU^T = I$ and $VV^T = I$) and $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_r)$ is a diagonal matrix with $\sigma_1 \geq \sigma_2 \geq \ldots \sigma_r \geq 0$ and $r = \min(m, n)$. LSA makes use of the matrix approximation theorem, which states that

$$\underset{W_k \text{ has rank } k}{\text{argmin}} \|W - W_k\|_2 = U_k \Sigma_k V_k^T \tag{4}$$

with U_k and V_k consist of the first k columns of U and V respectively and $\Sigma_k = \text{diag}(\sigma_1, \ldots, \sigma_k)$. So, for $k \leq r$ W_k is the best rank k approximation of W in the least square sense. The approximation error is bounded with respect to the Frobenius norm by $\|W - W_k\|_F \leq \sigma_{k+1}$. The column vectors of $(\Sigma_k V_k^T)$ are the new document vectors in the latent space. The mapping of some original document vector $\boldsymbol{d}$ into the latent space is described by $U^t\boldsymbol{d}$.

The SVD alters the original values in the matrix W by new estimates, based on the observed co-occurrences of terms and their 'true semantic meaning' within the whole corpus of documents [8]. The latter is achieved because terms with a common meaning are roughly mapped to the same direction in the latent space. By leaving out the smallest singular values, 'weak patterns' or noise are filtered out. The choice of k determines the degree of reduction, and it is therefore important to note that a high k value (corresponding to a weak reduction) might not be able to filter out noise or unimportant fluctuations in the source data, while a very small k value (strong reduction) will retain too little information from the original data structure [7].

2.3 Similarity score calculation and indexing of the HUPO PPP dataset

The HUPO PPP dataset was obtained from the PRIDE database[3], under accession numbers 4 to 98. These data sets are also accessible as PRIDE XML files via FTP[4].

All 95 Hupo PPP experiments and their corresponding peptides are directly taken from the PRIDE database and give a term/document matrix of the dimensions $25,052 \times 95$. Entries of the term/document matrix are weighted using IDF. In contrast to IR-applications on natural language, no further pruning of

[3] http://www.ebi.ac.uk/pride
[4] ftp://ftp.ebi.ac.uk/pub/databases/pride

unique terms was performed. A close look at the term/document matrix reveals the following differences compared to natural language datasets: while in a corpus of natural language text documents (for example the TREC Spanish AFP collection or the TREC Volume 3 corpus) the amount of unique words is below 2% [3], the Hupo PPP data set has 14,808 unique peptides (59%). This finding seems to contradict the intuitive assumption that proteomics experiments from the same tissue should yield highly similar results. The lack of reproducibility across proteomics experiments plays a considerable role in this divergence of the results [21]. This is illustrated for the HUPO PPP data by the fact that not even a single peptide is seen in every experiment. Moreover, only 37 peptides out of the 25,052 peptides are found in at least half of the experiments. In contrast, more than 70% of the reported proteins are only found in one or two experiments. This effect can be further explained by the wide array of techniques applied by the HUPO PPP contributors, with the express purpose of enhancing coverage and allowing subsequent method evaluation [19]. Furthermore, a shotgun proteomics approach to analyze a complex mixture typically results in approximately 30% of all proteins identified by only a single peptide [18].

As a result, non-zero entries constitute less than 4% of the document/term matrix, which, although better than a typical natural language set, is still quite poor, especially considering the fact that we analyzed a small experiment corpus and that this data set describes the proteome of a single tissue, namely plasma. In order to analyze the ability of LSA to compensate for this sparseness of the document/term matrix, the similarity of pair of the 95 experiments is computed using different values for k and the standard cosine measure, which gives $95(95-1)/2 = 4,465$ similarity scores. The choice of k depends on the distribution of the singular values of the original document/term matrix. A rapid drop of the values in the sorted sequence of singular values (i.e. $\sigma_l - \sigma_{l+1}$ is large and σ_{l+1} is small) indicates that W_l is good approximation with low error. In our case, we used $k = 75$ for small degree of compression and $k = 15$ for high compression. For comparison, similarity scores are also directly computed from the vector space representation.

3 RESULTS

As expected, the distribution of the similarity scores obtained for the HUPO PPP experiments with the VSM shows a low overall similarity (93.5% pairs of experiments have a similarity less or equal than 0.1).

Even the similarities for k=75 show no drastic improvements; the similarity scores rise only slightly (88.5% pairs of experiments have a similarity less or equal than 0.1). However, with a strong dimensionality reduction (k = 15) of the HUPO PPP data, latent semantic relationships between terms as well as co-occurrences of peptides within the replicate experiments are amplified. Consequently, the inter-experiment similarities rise, resulting in the fact that now only 35% (in contrast to 93.5% for the VSM) of the experiments pairs have a similarity of 0.1 or less. This effectively compensates for the sparseness of the matrix with all

entries are non-zero. To validate the results and show that LSA indeed resulted in meaningful experiment similarity, the obtained scores were visualized in a gray-scale map with white representing a score of 0, black a score of 1. The experiments are grouped by metadata, i.e. used technology (depletion step(s) applied; protein fractionation technique; search engine and finally mass spectrometer type). Since all 95 experiments originate from the same tissue, it is reasonable to expect LSA to amplify the intra-similarities within the same technology group.

3.1 Influence of Technologies in the Hupo PPP dataset

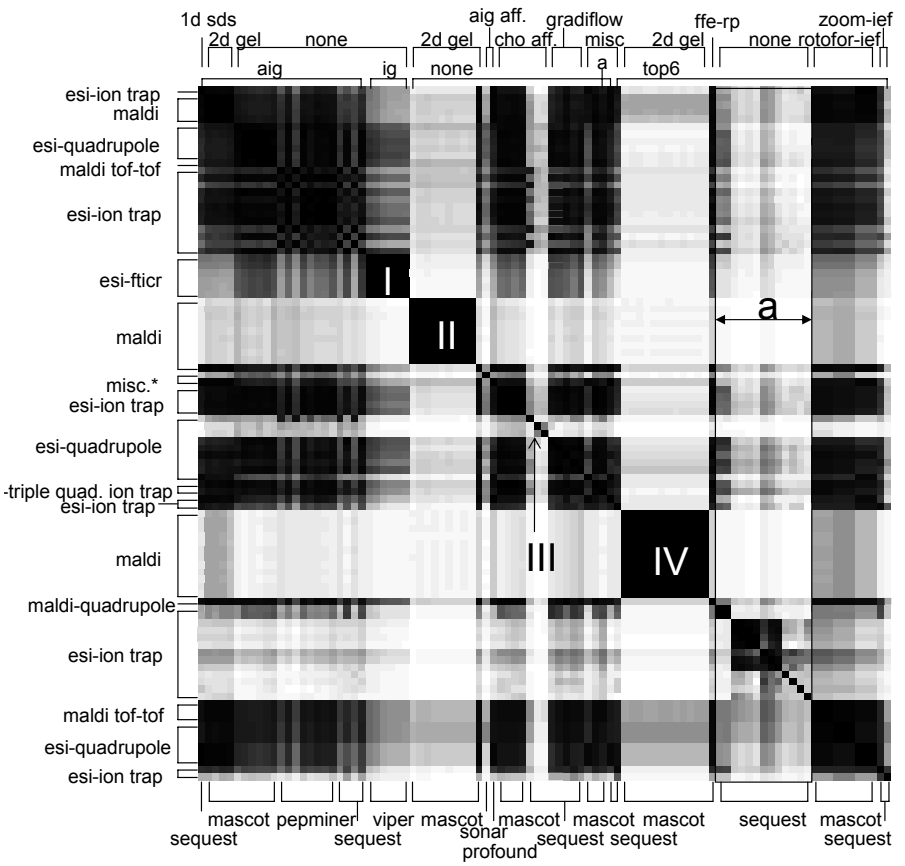

Fig. 1. A visualization of the inter-experiment HUPO PPP similarity matrix, obtained from an LSA with k =15. Dark indicates high similarity. The 95 experiments have been grouped by depletion technique, search engine, separation method and mass spectrometer (annotated above, to the left and below).

A gray-scale map of the similarity scores between all 95 experiments, obtained after LSA with $k = 15$, is shown in figure 1. The experiments have been grouped according to the four abovementioned technologies, and these have been annotated above, below and to the left-hand side of the map. Obviously, an experiment is identical to itself, which is why the top-left to lower-right diagonal is black. A great amount of experiments have a high similarity (dark areas) as expected when looking on one single tissue and using a low k, although some experiments are less similar to the other experiments than expected. Those experiments form distinct clusters, each with high internal similarity (I, II and IV). The first cluster (I) represents only ESI FT-ICR[5] experiments. This uniqueness in the instrument used, together with the use of proprietary VIPER search engine most probably contributes to the dissimilarity from the rest of the experiments. Cluster (II) is derived from a set of 2D-PAGE experiments, and these can be compared to cluster (IV), because both represent experiments performed by the same lab with the same technology. The only difference is the use of the top-6 protein depletion[6] on the biological sample in cluster (IV), while no depletion was employed in cluster (II). The very low similarity between these two clusters (nearly white overlap regions) shows that removal of the six most abundant proteins in plasma resulted in the detection of an almost completely different part of the plasma proteome. Three experiments (III) have very low similarity with the other experiments combined with a low similarity between each other. The reason for this could be the combination of a CHO-affinity (aldehyde affinity) fractionation and the SEQUEST search engine which no other experiment employed. As all three experiments are from the same laboratory (which only contributed these three experiments) and they have a rather low similarity among themselves, it seems plausible that these experiments are outliers and even might indicate suspect results. Another group of experiments also sticks out (band a). This group comprises experiments that employed a peptide shotgun approach (with no protein separation technique) on top-6 depleted samples. The shotgun experiments thus reveal very little similarity, both within the repeated experiments as well as compared to the rest of the experiments. The low similarities of the three clusters (I,II,IV) and the shotgun experiments (none for separation technique) with all other experiments indicates that they contribute unique peptide identifications (i.e.: they cannot be semantically connected to other peptide identifications). However, in contrast to the shotgun experiments, the 2D-PAGE cluster (II,IV) are strongly internally consistent, hinting at a high reproducibility of the method. The total number of peptide identifications for the HUPO PPP data set reveal that shotgun experiments contributed a major part of the overall unique peptide identifications. Finally, all observed clusters in this analysis derive from differences introduced by the various methodologies and technology platforms employed, rather than from differences between the samples which shows that a strong bias is introduced.

[5] EletroSpray Ionization Fourier-Transform Ion Cyclotron Resonance instrument
[6] The six most frequent, known proteins are removed.

3.2 Sample Analysis

A second experimental setup was used to evaluate the performance of LSA to compute meaningful similarities on proteomics data: instead of the natural grouping of peptides by experiment, all peptides found by any number of experiments of the same biological sample (plasma or serum) and anticoagulation treatment (EDTA, citrate or heparin) where selected and grouped, which resulted in a $5 \times 25,052$ document term matrix. Again, a VSM approach is not able to produce meaningful similarities, whereas LSA with k=2 (we would expect the document/term matrix to capture two semantic topics: plasma and serum) yields easily interpretable results. Those similarity scores are visualized and annotated in figure 2: VSM (on the left) only shows the (trivial) high similarities of one sample/anticoagulation group with itself, whereas LSA is able to resolve the similarity of the 4 groups of peptides originating from the plasma samples. Obviously the very low similarity between plasma and serum is caused by the fact that the serum samples do not contain any proteins (and therefore peptides) associated with clotting, such as Fibrin.

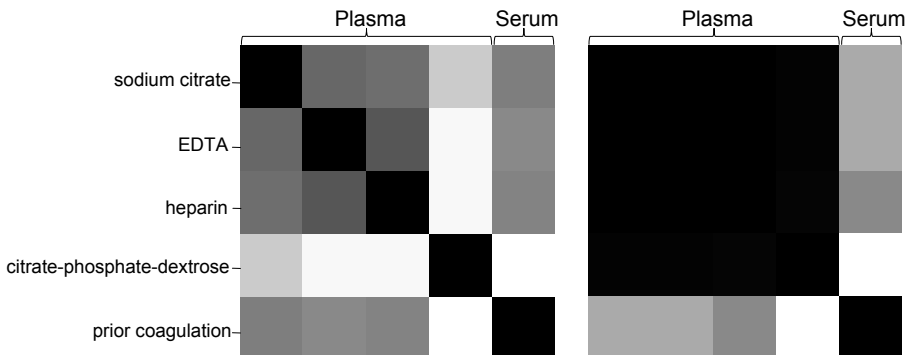

Fig. 2. : HUPO sample similarity matrices, for the VSM on the left and LSA with k =2 on the right. This grayscale map visualizes the differences in similarities resulting from a VSM and LSA (k=2) analysis on all peptides from the HUPO PPP dataset. Peptides found by any of the 95 experiments are grouped by the biological sample type (plasma or serum) - annotated above - and the anticoagulation method used - annotated to the left.

3.3 Interpretation of the peptide-based LSA

Literature suggests that a comparison based on exact matching of peptide sequences dramatically underestimates the overlap between experiments [17]. Therefore, it is particularly interesting to see how LSA groups peptides to latent topics, which are the dimensions of the latent space.

This analysis can be carried out by studying the term representation $U_k\Sigma_K$. A k-means clustering [23,24] performed on the rows of $U_k\Sigma_K$ allows detection of these related terms. Upon analysis of the resulting groups, two distinct patterns emerge. First of all, LSA resolves peptides distinguished only by the occurrence of isobaric amino acids (e.g.: isoleucine/leucine). Since these amino acids are indistinguishable to the mass spectrometer, their substitution does not affect the semantic representation of the containing peptide sequence. Second, peptides that represent subsequences of longer peptides, either through missed cleavages (e.g.: YLGNATAIFFLPDEGK and YLGNATAIFFLPDEGKLQHLENELT), in-source decay or in vivo proteolytic degradation (e.g.: YLGNATAIFFLPDEGK-LQHLENELT and YLGNATAIFFLPDEGKLQHLENELTHD) are all grouped together with the longer sequence. These two effects can be compared to synonyms in natural language.

4 Discussion/Conclusion

We have demonstrated a novel application of LSA by comparing peptide lists derived from many different proteomics experiments performed on the same tissue. By applying LSA to the data from the Hupo PPP study, we were able to show that this method can handle the very diverse and heterogeneous data arising from proteomics experiments and compute meaningful similarities.

A large amount of experiments have a high similarity after LSA, which shows the strength of the method. However, some experimental setups, namely 2D-PAGE and shotgun approaches, strongly bias the set of observed peptides, which LSA cannot compensate for. Our results confirm visually, that if the goal of a project is to achieve maximal proteome coverage for a particular sample, shotgun proteomics experiments, repeated over multiple replicates achieve the most gain. 2D-PAGE analysis should not be disregarded as an analytical tool however, since it can complement a substantial fraction of unique identifications. Due to the high internal reproducibility of 2D-PAGE analyses as performed in the HUPO PPP, it seems that carrying out many replicates of this technology does not necessarily lead to a proportional increase in novel peptide identifications. In the specific case of plasma, the influence of various depletion techniques is also of interest. While methods employing top-6 depletion contributed more than 50% of the identifications, about 10% of all proteins were only found when no depletion was used at all.

The relatively simple task of comparing different sample types demonstrates that the fundamental difficulties arising from the origin of the data could be overcome through the utilization of an LSA analysis and its key principle of peptide/experiment association data representation in a lower dimensional 'latent space'. It is important to consider that the latent semantic analysis employed here greatly benefits from the large number of varying experimental repetitions on the same sample.

4.1 Interpretation of the semantic associations

An interesting finding is the ability of LSA to detect semantic relationships between apparently unrelated sets of peptide sequences, based solely on co-occurrences within experiments. We have found that at least some of these semantic links can be explained by underlying methodological or biological concepts and can be compared to synonyms found in natural language. The application of LSA to replicated shotgun experiments might help to alleviate one of the primary caveats of peptide-centric proteomics: the protein inference problem.

Since the semantic structures underlying protein lists (at least in part) represent entities of biological interest, the nature of the semantic relationships that occur at that level are also of considerable interest. Potential candidates of biological importance include protein complexes or protein components of the same pathway.

4.2 Future perspectives

It is clear from these findings that large collections of heterogeneous proteomics datasets can be mined relatively easily to obtain valuable information with LSA. The analysis carried out opens many paths for further investigations. By extending the analysis to include other tissue data sets (for instance the HUPO Brain Proteome Project (HUPO BPP) [9], and eventually any available proteomics data) and by carefully choosing an appropriate value for k, the focus of investigation could be shifted from the fine-grained effects resulting from the application of different technology platforms, to the course-grained distinctions derived from differences in tissue type, disease state or developmental stage.

It is of significant interest to get a better understanding of the semantic similarities peptide share in the latent semantic space resulting from singular value decomposition. Other methods, especially 'probabilistic latent semantic indexing' described in [11] which has a solid statistical foundation should be examined.

5 Acknowledgements

This work has been supported by the EU "ProDaC" grant number LSHG-CT-2006-036814 and the BBSRC "ISPIDER" grant.

References

1. R. Aebersold and M. Mann. Mass spectrometry-based proteomics. *Nature*, 422:198–207, 2003.
2. R. Craig, J. P. Cortens, and R. C. Beavis. Open source system for analyzing, validating, and storing protein identification data. *J Proteome Res*, 3:1234–1242, 2004.
3. M. W. Davis and W. C. Ogden. Free resources and advanced alignment for cross-language text retrieval. *TREC*, pages 385–395, 1997.

4. S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.
5. F. Desiere, E. W. Deutsch, A. I. Nesvizhskii, and P. Mallick. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol*, 6:R9, 2005.
6. B. Domon and R. Aebersold. Mass spectrometry and protein analysis. *Science*, 312:212–217, 2006.
7. S. T. Dumais. Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments and Computers*, 23:229–236, 1991.
8. G. W. Furnas, S. Deerwester, S. Dumais, and T. K. Landauer. Information retrieval using a singular value decomposition model of latent semantic structure. *Proc. of SIGIR'88 Conference on Information Retrieval*.
9. M. Hamacher, R. Apweiler, G. Arnold, and A. Becker. Hupo brain proteome project: summary of the pilot phase and introduction of a comprehensive data reprocessing strategy. *Proteomics*, 6:4890–4898, 2006.
10. H. Hermjakob and R. Apweiler. The proteomics identifications database (pride) and the proteomexchange consortium: making proteomics data accessible. *Expert Rev Proteomics*, 3:1–3, 2006.
11. T. Hofmann. Probabilistic latent semantic indexing. *Proc. SIGIR'99 Conference on Information Retrieval*.
12. P. Jones, R. G. Cote, L. Martens, and A. F. Quinn. Pride: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Res*, 34:D659–663, 2006.
13. S. M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Trans. Acoustics, Speech and Signal Processing*, 35:400–401, 1987.
14. T. K. Landauer, P. W. Foltz, and D. Laham. Introduction to latent semantic analysis. *Discourse Processes*, 25:259–284, 1998.
15. Mallick, Schirle, Chen, and Flory. Computational prediction of proteotypic peptides for quantitative proteomics. *Nat Biotechnol*, 25:125–131, 2007.
16. L. Martens, H. Hermjakob, P. Jones, and M. Adamski. Pride: the proteomics identifications database. *Proteomics*, 5:3537–3545, 2005.
17. L. Martens, M. Muller, C. Stephan, and M. Hamacher. A comparison of the hupo brain proteome project pilot with other proteomics studies. *Proteomics*, 6:5076–5086, 2006.
18. A. I. Nesvizhskii and R. Aebersold. Interpretation of shotgun proteomic data: The protein inference problem. *Mol Cell Proteomics*, 4:1419–1440, 2005.
19. G. S. Omenn, D. J. States, M. Adamski, and T. W. Blackwell. Overview of the hupo plasma proteome project. *Proteomics*, 5:3226–3245, 2005.
20. J. Prince, M. W. Carlson, R. Wang, P. Lu, and E. M. Marcotte. The need for a public proteomics repository. *Nat Biotechnol*, 22:471–472, 2004.
21. K. A. Reidegeld, M. Muller, C. Stephan, and M. Bluggel. Abstract the power of cooperative investigation: summary and comparison of the hupo brain proteome project pilot study results. *Proteomics*, 6:4997–5014, 2006.
22. G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the Acm*, 18:613–620, 1975.
23. A. Smellie. Accelerated k-means clustering in metric spaces. *Journal of Chemical Information and Computer Sciences*, 44:1929–1935, 2004.
24. D. Steinley. K-means clustering: A half-century synthesis. *British Journal of Mathematical & Statistical Psychology*, 59:1–34, 2006.