Journal of
CHEMICAL INFORMATION
AND MODELING

Article

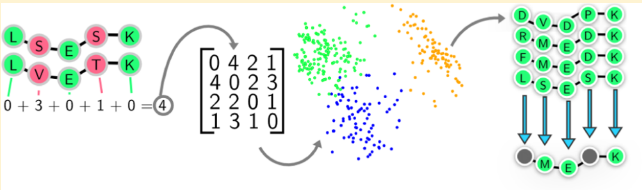
pubs.acs.org/jcim

Standardizing and Simplifying Analysis of Peptide Library Data

Andrew D. White, Andrew J. Keefe, Ann K. Nowinski, Qing Shao, Kyle Caldwell, and Shaoyi Jiang*

Department of Chemical Engineering, University of Washington, Seattle, Washington, United States

Supporting Information



ABSTRACT: Peptide libraries allow researchers to quickly find hundreds of peptide sequences with a desired property. Currently, the large amount of data generated from peptide libraries is analyzed by hand, where researchers search for repeating patterns in the peptide sequences. Such patterns are called motifs. In this work, we describe a set of algorithms which allow quick, efficient, and standard analysis of peptide libraries. Four main techniques are described: (1) choice of the number of motifs present in a peptide library; (2) separation of the peptides into groups of similar sequences; (3) fitting of a model to the peptides to extract motifs; (4) analysis of the library using quantitative structure−property relationships if no clear motifs are present. The application of five previously published data sets shows these techniques can automatically repeat the work of experts quickly and allow much more flexibility in analysis. A new way of visually presenting peptide libraries is also described, which allows visual inspection of the grouping and spread of sequences. The algorithms have been implemented in an open-source plug-in called "peplib" and an online web application.

■ INTRODUCTION

Combinatorial peptide libraries are powerful tools for quickly screening millions of peptides for activity. With an appropriate assay, it is possible to obtain the individual sequences of active peptides. This may be used to discover protein ligands,[1] antimicrobial peptides,[2] and even molecules for protein separation.[3] Peptide libraries are a large collection of peptides each with different sequences. A process is applied to the library that separates active peptides from inactive peptides. For example, running the library over a column with an immobilized target molecule will elute the nonbinding peptides away from those which bind. Then, the peptides which are bound may be examined to identify active sequences. The remarkable aspect of peptide libraries is that millions of sequences can be tested in parallel, enabling high-throughput experiments.

One of the most accurate types of peptide libraries are solid-phase combinatorial libraries.[4,5] Solid-phase libraries are unique in their ability to eliminate biases in amino acid frequency while still providing individual active sequences. Some peptide library methods have confounding factors; for example, FLITRX libraries, which display peptides using *E. coli*, are estimated to lose around 10% of peptides due to expression problems.[6] Bacteria or phage libraries require multiple iterations and tuning to ensure that multiple active sequences are discovered.[7] Other methods can have convoluted results. For example, using affinity columns with the target bound to the column and peptides in the mobile phase screens for both peptide abundance and affinity,[8] whereas peptide affinity is the only variable with which we are concerned. Solid-phase libraries are well suited to analysis because they provide individual sequences based only on affinity.[4] Utilizing solid-phase libraries consists of three basic steps: synthesis, activity determination, and sequencing. During synthesis, the library of peptides is constructed on solid particles, ranging from 90 up to 200 μm depending on the chemistry and application.[5,9] Each particle, or bead, contains tethered peptides, all with the same sequence. Between beads, however, there are different sequences. During the activity determination step, each bead is tested and beads which are active are isolated. This is done, for example, using a colorimetric or fluorescence assay, and the beads may be separated via an automated sorter.[10] Finally, during sequencing, the peptides are cleaved one amino acid at a time and sequenced using MALDI-TOF in a technique called partial edman degradation.[4] The result of these steps is a list of sequences which are active.

There are two techniques used for analyzing the peptide libraries in this work. The first is quantitative structure−activity relationships (QSAR), which excel at describing small molecules. A descriptor is a quantitative metric based on chemical structure, for example number of double bonds, which may then be correlated with activity. The correlation of a descriptor with activity is called a QSAR. The second technique is motif discovery. Motifs are frequently occurring short strings of amino acids. For example, the three amino sequence "TYG" could be a motif, and the one letter amino acid abbreviations are used. Typically, motif lengths are between 3 and 10 amino acids long.

Analysis of these solid-phase peptide libraries is still relatively unexplored, excepting traditional consensus sequence analysis. There are three main challenges for analysis of such experiments. The first is that the variable regions of the peptides are typically too short (3−10 amino acids) to be

Received: October 10, 2012
Published: January 17, 2013

© 2013 American Chemical Society

analyzed using existing techniques from proteomics or genomics. For example, the popular Multiplied EM for Motif Elicitation algorithm, which discovers sequence motifs, is suggested to work on at least 8 amino acid length peptides and typically used for searching whole proteins or long gene sequences for motifs.[11,12] The second challenge is that, when viewed from a traditional QSAR descriptor based perspective, the peptides have molecular weights far beyond what most descriptors were designed for. This limits the applicability of QSAR techniques, and even analysis of two amino acid dipeptides is challenging.[13] Finally, the results of peptide libraries are a list of active sequences. This makes it difficult to utilize the large number of QSAR classification techniques which require examples of both active and inactive structures.[14] Experimentally, it is possible to isolate and identify inactive sequences, but the results will be quite close to random peptide sequences, providing little information. These challenges limit the use of the large amount of data generated from such experiments.

In this work, we describe a collection of algorithms meant to solve these challenges and simplify the analysis of peptide library data. Most of the algorithms operate on the sequence view of the peptides; each peptide is represented as a string of letters. This is the most relevant perspective in biology, where a consensus motif or sequence is the desired output from a peptide library. It is often the case, however, that certain active sequences do not contain a consensus motif and thus we also describe algorithms which examine peptides from a molecular perspective using traditional QSAR descriptors.

The algorithms produce four important results. The first is the number of motifs present in a peptide library. The second is the grouping of the sequences based on the number of motifs. Although not discussed in this work, the grouping is enough to produce a substitution matrix for PSI-BLAST to find examples of the grouped sequences in protein databases.[15] The third result is the motifs of the grouped sequences, as determined from a model fitting procedure. These motifs are ultimately the output of a peptide library. The motifs indicate the preferred substrate for the peptide library assay. Finally, if the motif fit is unsatisfactory, QSARs may be calculated to test if other structure−property relationships fit the sequences better than motifs.

■ METHODS

An overview of the techniques of this work is given in Figure 1. Sequence−sequence distances are calculated, a distance matrix may be derived from these distances, the matrix is used to separate the sequences into clusters, and motifs are fitted to each cluster. Separately, QSAR descriptors are calculated on all sequences and their distributions are different than that of the entire peptide library.

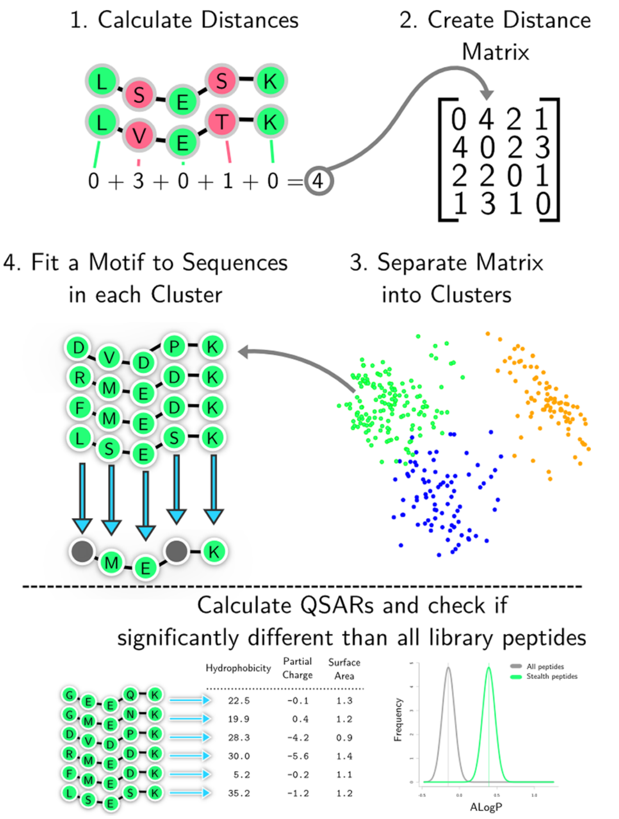


Figure 1. Overview of the techniques presented in this work for analyzing peptide libraries. Sequence−sequence distances are calculated (1), a distance matrix is derived from these distances (2), the matrix is used to separate the sequences into clusters (3), and motifs are fitted to each cluster (4). Separately, QSARs are calculated on all sequences and descriptors which are significant relative to the inactive sequences are identified.

Generally multiple sequences will be active in a peptide library. If most of these sequences are similar, they may be considered to contain a common sequence pattern or motif. The motif may be determined by an expert examining the collection of active sequences. In some cases, there may even be multiple unrelated motifs. It is difficult to effectively and objectively categorize the sequences into a small number of motifs and also choose the number of motifs that are being observed. This is also complicated due to the large number of active sequences generated from solid-phase peptide libraries (100−500). In machine learning, grouping sequences together is called clustering and is a routine problem. If the number of motifs is known in advance, it is possible to collect the sequences into groups based on their chemical similarity.

K-means is a clustering technique that groups points, or sequences, into groups based on their similarity.[16] K-means clustering performs operations on a symmetric $N \times N$ distance matrix, where element i,j encodes the distance from sequence i to j. Although it is possible to consider variable length peptide libraries using multiple sequence alignment tools,[17] we will assume that each sequence has the same length within each peptide library. Multiple sequence alignment may also address any frame-shift issues, where motifs are shifted in position. This is rare in peptide libraries where motif position is important due to the short sequence length and was not observed in the data sets considered here. The distance between two sequences may be calculated using substitution scores according to

$$d_{ij} = \sum_{k=1}^{l} f(s_{ik}, s_{jk}) \tag{1}$$

Where, l is the length of the peptide, $f()$ is the substitution function (for example, the BLOSUM62 matrix[18]) which measures how chemically similar two amino acids are, and s_{ik} is the kth position of sequence i. If a BLOSUM matrix is utilized for the substitution function, the distance matrix must be shifted. Positive numbers indicate similarity in BLOSUM and low negative numbers indicate dissimilarity. Thus, the shift should make the most positive distance be the zero element and the most negative element become as large as the largest difference in the unshifted distance matrix. BLOSUM85 is used in this work, based on a comparison of BLOSUM50, BLOSUM62, BLOSUM85, BLOSUM90, and the Hamming distance in the Supporting Information.

K-means clustering is a "hard" clustering technique, meaning that each sequence may belong to only one cluster. The Hartigan and Wong algorithm[19] was used as implemented in the statistical computer program R[20] with 20 random starts to account for the local optimization. The clustering with the minimum sum of the square distance to the cluster means was chosen among the 20.

Agglomerative, or hierarchical, clustering techniques are commonly used for clustering genetic sequences.[21] To compare with the K-means clustering, single-linkage agglomerative clustering was examined in the Supporting Information and found to provide worse results in 8 out of the 10 comparisons with K-means.

The distance matrix constructed using a substitution function may be visualized using a principal component analysis in order to visualize the results from the K-means clustering. Principal component plots were calculated by first computing the eigenvector matrix of the sequence−sequence distance matrix using singular value decomposition and then multiplying the data matrix by the eigenvector matrix. Only the first two columns are retained, which is the data projected onto the principal two components. The implementation of this algorithm in R was used.[20]

Choosing the number of clusters, and thereby the number of motifs, is an ongoing research problem. One way to choose the number is the "elbow" technique, which is utilized here.[22] In the elbow technique, some measure of the goodness-of-fit as a function of the number of clusters is plotted. The goodness-of-fit generally increases as a function of the number of clusters; more model parameters create a better fit. The number of clusters just before the goodness-of-fit flattens may be chosen as the number of clusters. This involves searching for an elbow in a plot, hence the name "elbow" technique.

The choice of the number of clusters should integrate information known by researchers as well, since ultimately the correctness of a clustering depends on the data set and experiment. For example, the elbow technique may be used to create a hypothesized grouping of a set of sequences into a few clusters. Next, representative peptides from each cluster may be synthesized and tested to see if there are true chemical differences between the clusters. Such an approach was utilized in ref 4. The number of clusters should be thought of as an exploratory variable which must be justified at the end of analysis.

The clustering algorithms described above assign each sequence to a particular cluster. The next step in the analysis is to extract the motif or consensus sequence. Here, we use a modified version of MEME.[11,12,23] MEME is an expectation-maximization (EM) algorithm that identifies the motif in a set of sequences and the location of the motif in the sequences. The algorithm optimizes the likelihood of a particular motif, which is a measure of how well the proposed motif fits the data. The likelihood for a collection of n sequences, each of length l, given model M is given by

$$L(S;\, M) = \prod_{i=1}^{N} \prod_{z=1}^{l-w} [\prod_{j=0}^{z-1} \Pr(s_{ij}|b)][\prod_{j=z}^{z+w} \Pr(s_{ij}|m_j)] \\ [\prod_{j=z+1}^{w} \Pr(s_{ij}|b)] \tag{2}$$

Where s_{ij} is the jth position of the ith sequence, b is the background model, which models the nonmotif positions, m_j is the jth position of the motif, N is the number of sequences, l is the length of the sequences, w is the length of the motifs, and z is the starting location of the motifs. Each probability, $\Pr(\cdot|\cdot)$, is a normalized vector of probabilities, with one probability for each possible amino acid occurring. Following the general EM algorithm, this equation is not maximized directly, but instead the expectation of its log over the possible starting positions (z) is maximized. Details of general EM algorithms may be found in Dempster et al.[24] The "hidden" data, a component of applying the EM algorithm, is the motif starting position, z. The motif model equations produce a $w \times A$ matrix, where w is the width of the motif and A is the number of possible amino acids (typically 20). Each column in the matrix represents the probabilities of each amino acid at that motif position. If the motif position is not variable, then the model reduces to the proportion of each amino acid at the motif positions. This is the usual quantity analyzed for determining motifs from a collection of sequences. See Sweeney et al.[4] or Chen et al.[25] for examples. The expected log-likelihood, the maximized quantity in fitting the model, is the goodness-of-fit used above in the elbow technique.

In the implementation of the MEME algorithm in ref 23, initial guesses based on statistics derived from genomic analysis were used; the frequency of amino acids is nonuniform in naturally occurring proteins.[11] However, in our implementation we use uniform distributions throughout for initial guesses due to the uniformity of amino acid representation in solid-phase peptide libraries. The equations for this method are shown in Supporting Information eqs S1−S3. The EM algorithm is considered converged here once the sum of the squares of the model parameters of the model changes by less than 0.1%.

The width of the motif model is chosen a priori and is ultimately a decision of the expert using the algorithms. The procedure used in this paper which recovered motifs seen by experts analyzing the data sets is to start with a motif width of three and increase by one amino acid as long as the additional motif positions have one amino acid with greater than 15% probability mass. Another method is to create elbow plots of the log-likelihood as a function of motif width.[22] The starting position, z, of the motifs may be set as to be the same for all sequences or different for each motif. If it is the same for all sequences, the EM algorithm operates on z as a vector and if it is different for each sequence a matrix is used for z. Throughout this work, it is not assumed that the motif starts at the same position for each sequence, and hence a matrix is used.

A complementary approach to motif searching is to examine QSARs.[26] QSAR descriptors are functions which take a peptide sequence as an input and output a number representing some property of the peptide. For example, the length of a peptide sequence is a QSAR descriptor. A QSAR is a function which converts these descriptors into a quantitative estimate for the activity of a sequence. QSARs may be more appropriate for analyzing peptide libraries when the chemical properties of the peptides seem more important than the sequence. Searching for

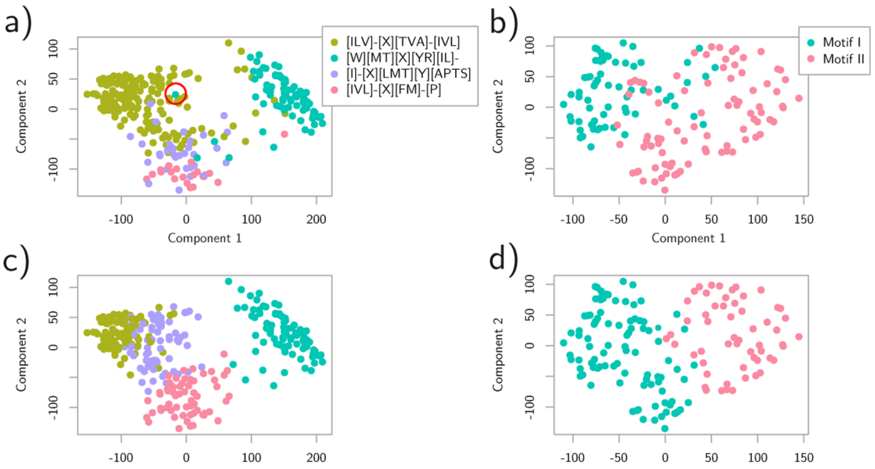


Figure 2. Plot of the two principal components of the distance matrix of the peptide library data from two previously published data sets.[4,25] The colors in a and b represent the segregation of sequences according to Sweeney et al.[4] and Chen et al.,[25] respectively. The consensus sequences developed by Sweeney et al.[4] are in panel a in the legend. No consensus sequences were reported for panel b but they were separated by Chen et al.[25] into two motifs. Square brackets denote the motif positions, and "-" indicates a nonmotif, or background, position. The amino acids in brackets are ordered by most frequently occurring to least in the motif positions. The results from K-means clustering are shown in panels c and d. The K-means clustering produces similar results, capturing the key features of the sequences. The point circled in red in panel a is a possible misclassification. These data sets were chosen because they have been previously clustered in Sweeney et al.[4] and Chen et al.[25] The proportion of variance explained in these principal component analysis (PCA) projections is 0.64 in panel a and 0.56 in panel b.

a motif is relatively simple, because if a pattern appears multiple times in active sequences then it is likely significant (assuming a uniform background distribution of amino acids). However, a QSAR descriptor may contain the same value for each active sequence simply because all peptides have that property. A trivial example would be the number of chlorine atoms being zero for all active sequences. There are no amino acids with chlorine atoms and thus all peptides in the library would have the same QSAR descriptor value. A more subtle example is the surface area of the molecules, which is similar because it is mostly a function of the length of the peptides and all peptides are generally the same length in a peptide library. Thus, it is important to calculate QSAR descriptors on both the active sequences and the inactive sequences. If the distributions of a QSAR descriptor are significantly different, then it may be said to be a relevant QSAR descriptor. This significance may be calculated by using a Wilcoxon test,[27] a nonparametric version of the Student's t test. This hypothesis test is used to compare the median of two distributions and results in a p-value. A low p-value means the distributions are different and that the QSAR descriptor is significant. The only complication is calculating a QSAR descriptor on the distribution of inactive sequences, which is unknown. One approximation that may be easily checked during the assaying phase of peptide libraries is that the active sequences are a small fraction of the peptide library. If that is the case, and due to the lack of bias in solid-phase peptide library synthesis, then we may estimate the inactive sequences as uniformly random sequences. These may be randomly generated computationally to construct a distribution of QSAR descriptors on the inactive sequences for the Wilcoxon test.

The Wilcoxon paired signed rank test (Wilcoxon t test) was used as implemented in R.[27] In order to estimate the QSAR descriptors on inactive sequences, 500 sequences were randomly generated assuming a uniform distribution of amino acids. QSAR descriptors were calculated on the randomly generated sequences for the Wilcoxon t test.

The QSAR descriptors reported in this research are intentionally simple and meant to illustrate the methods. The counts of groups were calculated as follows: the basic groups are H, R, and K. The acid groups are E and D. The aromatic groups are W, Y, and F. The polar groups are S, T, C, P, N, Q, K, R, H, E, and D. The charged groups are E, D, H, K, and R. ALogP is the average AlogP of the amino acids in the sequence, as calculated according to Ghose and Crippen[28] and implemented in the Chemistry Development Kit.[29]

To summarize the techniques presented, analysis of peptide libraries should begin with elbow plots to estimate the number of motifs present in a data set. Next, the sequences should be clustered and a motif model be fit to each cluster to describe the motifs. If the motifs have many variable positions, then QSAR descriptors may be fit instead to the sequences.

■ RESULTS

Five previously published data sets are used to test the analysis techniques presented. The first three come from Sweeney et al.,[4] and the second two come from Chen et al.[25] All five peptide libraries target different phosphatase enzymes and are solid-phase peptide libraries. The algorithms presented are general and do not require that the data come from solid-phase peptide library techniques[4] or phosphatase enzymes.

The principal component plots of two previously published peptide libraries are shown in Figure 2. The colors in Figure 2a and b represent the clustering of the sequences as done by the authors (experts) and not an algorithm. These two data sets were chosen because they were also clustered by experts. The distance matrix provides separation between the clusters which were chosen by experts. Some possible mistakes in the classification also become visible as well. For example, sequence 133 which is circled in Figure 2a, begins with isoleucine yet was classified into a cluster where each other sequence begins with tryptophan. The results using the K-means clustering are shown in bottom two panels in Figure 2c and d. The k-means clustering in Figure 2c and a have an 85% overlap, and b and d have 67% overlap. The clustering finds similar patterns and is

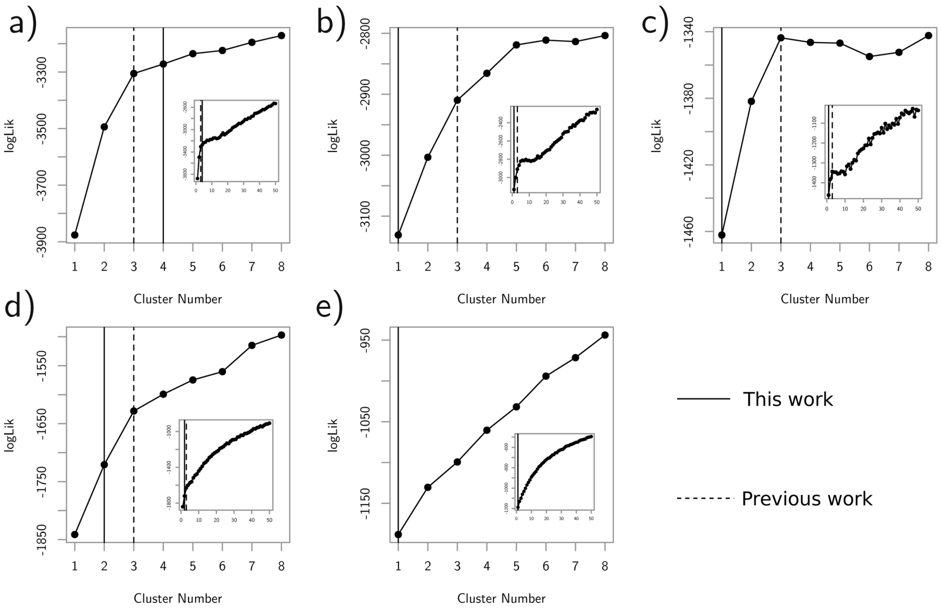


Figure 3. Elbow plots of the number of clusters or motifs in the five previously published data sets.[4,25] The *y*-axis is the log-likelihood of the motif models over the data set and the *x*-axis is the number of clusters. The dashed vertical line indicates where we choose the number of clusters, based on the elbow technique. The insets show the same plot but to a larger number of clusters. The elbows are more visible in these plots. The solid vertical lines show the choice of the experts who originally analyzed the data sets. Panels b, c, and e were not considered to have multiple motifs by experts. Panel e is not considered to have multiple motifs based on information described in text. These plots provide a justifiable and straightforward technique for choosing the number of motifs.

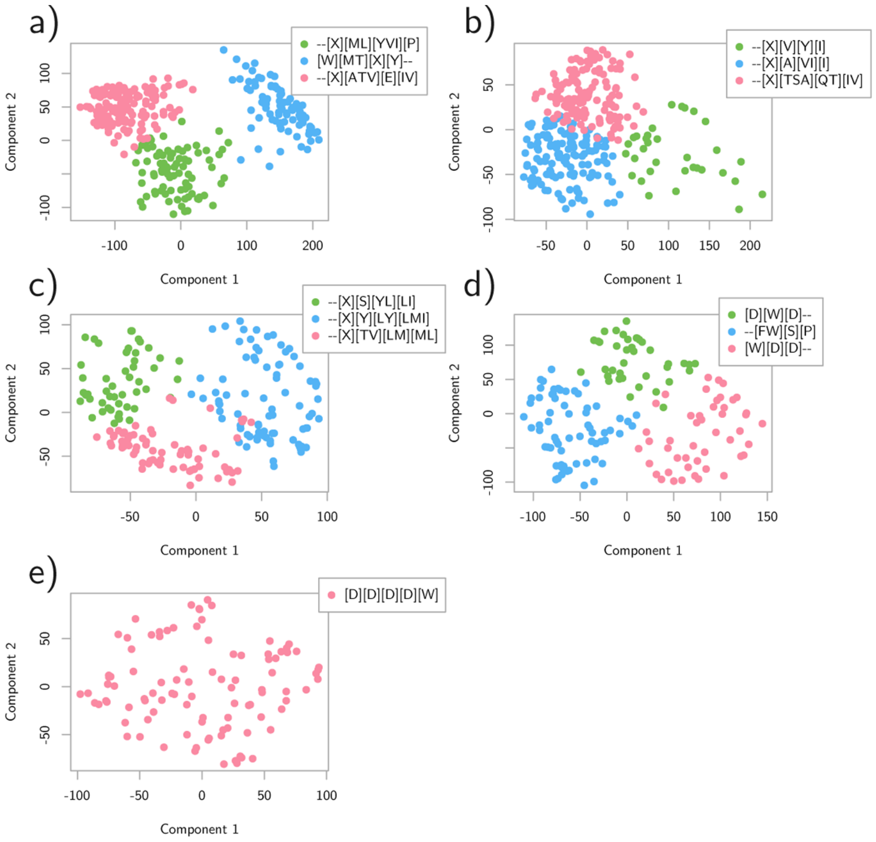


Figure 4. Plot of the two principal components of the distance matrix of the peptide library data from five previously published data sets. The colors correspond to the motifs shown in the legend. Square brackets denote the motif positions, and "-" indicates a nonmotif, or background, position. The amino acids in brackets are ordered by most frequently occurring to least in the motif positions. The numbers of motifs were chosen according to Figure 3. Panels a and c may be compared directly to experts' clustering and motif choices shown in Figure 2. The proportion of variance explained in these PCA projections in order from a–e is 0.64, 0.43, 0.52, 0.56, and 0.47.

automatic, reducing the risk of accidentally misclassifying a sequence. The K-means algorithm operates on the entire distance matrix, yet only the principal two components are shown in Figure 2, thus overlapping clusters there do not necessarily mean they are overlapping in the other dimensions which are not shown.

Elbow plots for the five data sets were used to determine the number of clusters present. These are shown in Figure 3. The solid vertical line indicates the number of clusters as determined from experts. The dashed vertical line indicates the number of clusters as determined from the elbow technique. In Figure 3a, the technique agrees well with the choice of experts, differing by one. In Figure 3b and c, the authors did not consider multiple motifs. In Figure 3d the technique disagrees with the choice of the experts by one motif. Only one cluster was chosen for Figure 3e because increasing the cluster number did not significantly change the motifs, which is another test to determine the number of motifs. Although the process is in some ways still subjective with elbow plots, they do provide an easily communicated and justifiable method to choose the number of motifs.

The results of the motif models on the 5 data sets are shown in Figure 4. Figure 4a may be compared with the clustering as accomplished by experts and their motif choices in Figure 2a. Although there are some differences, the clustering has an 85% overlap if the number of motifs are set to be four. The most significant partition is between the blue and other clusters. This is captured in both Figures 4a and 2a. The advantage of the technique presented here is that the process takes a few seconds and the effect of changing the number of motifs and motif width may be tested just as quickly. Figure 4b shows the effect of segregating the sequences into clusters in contrast to what is presented in Sweeney et al.,[4] where there is no separation. The motif model determined the last three residues to be the most important, with the blue points being the most conserved motif. The tyrosine in position 5 is unique in that cluster. In Figure 4c, it is possible to see a significant difference between the three motifs in positions two and three, where the green cluster shows a hydrophilic serine in position two and the blue cluster shows an aromatic tyrosine. The red cluster shows a threonine (similar to the serine in the green cluster) in position two but no aromatics in the third position.

Figure 4e, which shows the TULA-2 pre data set, is a good example of when QSAR descriptors are more appropriate than motifs. The motif in Figure 4e is four aspartic acids and a tryptophan. However, that particular motif actually does not appear in the list of active sequences. Most of the active sequences contain two acids and one aromatic group. A small set of QSAR descriptors on the active sequences were calculated along with their *p*-values, which are shown in Table 1 under the TULA-2 pre column. It is clear that the number of aromatic groups and acid groups is a significant QSAR descriptor. For comparison, the *p*-values are shown for the same analysis on the TULA-2 post peptide library from the same publication. That data set is shown in Figure 4d as well. The number of acidic groups is no longer a significant QSAR descriptor, but the number of aromatic groups still is. This is corroborated by the motif shown in Figure 4d, where each of the motifs contains an aromatic amino acid but not necessarily an acidic amino acid. QSAR analysis provides a complementary technique for finding patterns in peptide libraries when there is no clear motif.

Table 1. QSAR Descriptors from Two Peptide Library Data Sets[a]

QSAR descriptor	*p*-value TULA-2 pre	*p*-value TULA-2 post
basic group count	6.71×10^{-4}	2.88×10^{-7}
acid group count	8.71×10^{-2}	7.77×10^{-22}
polar group count	2.77×10^{-15}	4.57×10^{-3}
aromatic group count	2.65×10^{-48}	4.60×10^{-29}
charged group count	1.11×10^{-1}	2.22×10^{-5}
ALogP	7.38×10^{-9}	1.83×10^{-10}

[a] A lower *p*-value indicates a significant QSAR descriptor, as determined from a Wilcoxon T-test. A significant QSAR descriptor means that the sequences which were active in the library have a QSAR descriptor value significantly different than what was seen in the inactive sequences.

The algorithms presented here, along with additional analysis techniques for analyzing peptide library data, have been packaged into a plug-in for R called "peplib." It is available on the CRAN repository (http://cran.r-project.org) along with a manual describing its use. Further information and a web-based application which allows users to apply a simplified version of the algorithms in this paper may be found at http://peplib.org.

■ CONCLUSIONS

Solid-phase peptide libraries are powerful experimental techniques for quickly screening millions of peptides for activity. Analysis of such libraries is generally accomplished by experts analyzing hundreds of sequences by hand and using intuition for the number of motifs and consensus sequences. Reliable and freely available algorithms have been described here to analyze the data generated from such experiments quickly and automatically. We have described how to choose the number of motifs in a peptide library, how to group sequences together based on similarity, how to extract the motifs from similar sequences, and how to analyze the chemical properties of the peptides with QSARs. The algorithms compare well with the work of experts in the field on five previously published data sets and excel in their speed and consistency compared with the current techniques. Implementations of these algorithms, documentation, and a tutorial for them may be obtained at http://cran.r-project.org/web/packages/peplib/. An online application is also available at http://peplib.org. It allows researchers to use a basic version of the algorithms presented here on their own data.

■ ASSOCIATED CONTENT

S Supporting Information

Equations for the update steps of the MEME algorithm used in this work and comparison of BLOSUM matrices and clustering algorithms. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: sjiang@u.washington.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the Office of Naval Research (N00014-10-1-0600) and the National Science Foundation (CBET-0854298).

■ REFERENCES

(1) Obata, T.; Yaffe, M. B.; Leparc, G. G.; Piro, E. T.; Maegawa, H.; Kashiwagi, A.; Kikkawa, R.; Cantley, L. C. Peptide and protein library screening defines optimal substrate motifs for AKT/PKB. *J. Biol. Chem.* **2000**, *275*, 36108−36115.

(2) Blondelle, S. E.; Lohner, K. Combinatorial libraries: a tool to design antimicrobial and antifungal peptide analogues having lytic specificities for structure-activity relationship studies. *Biopolymers* **2000**, *55*, 74−87.

(3) Sennels, L.; Salek, M.; Lomas, L.; Boschetti, E.; Righetti, P. G.; Rappsilber, J. Proteomic Analysis of Human Blood Serum Using Peptide Library Beads. *J. Proteome. Res.* **2007**, *6*, 4055−4062.

(4) Sweeney, M. C.; Wavreille, A.-S. S.; Park, J.; Butchar, J. P.; Tridandapani, S.; Pei, D. Decoding protein-protein interactions through combinatorial chemistry: sequence specificity of SHP-1, SHP-2, and SHIP SH2 domains. *Biochemistry* **2005**, *44*, 14932−14947.

(5) Keefe, A. J.; Caldwell, K.; Nowinski, A. K.; White, A. D.; Thakkar, A.; Jiang, S. Screening Nonspecific Interactions of Peptides with Elminiated Background Interference. *Biomaterials* **2013**, *34*, 1871−1877.

(6) Lu, Z.; Murray, K. S.; Cleave, V. V.; LaVallie, E. R.; Stahl, M. L.; McCoy, J. M. Expression of Thioredoxin Random Peptide Libraries on the Escherichia coli Cell Surface as Functional Fusions to Flagellin: A System Designed for Exploring Protein-Protein Interactions. *Nat. Biotechnol.* **1995**, *13*, 366−372.

(7) Smith, G. P.; Petrenko, V. A. Phage Display. *Chem. Rev.* **1997**, *97*, 391−410.

(8) Songyang, Z.; Shoelson, S. E.; Chaudhuri, M.; Gish, G.; Pawson, T.; Haser, W. G.; King, F.; Roberts, T.; Ratnofsky, S.; Lechleider, R. J. SH2 domains recognize specific phosphopeptide sequences. *Cell* **1993**, *72*, 767−778.

(9) Yao, N.; Xiao, W.; Wang, X.; Marik, J.; Park, S. H. H.; Takada, Y.; Lam, K. S. Discovery of targeting ligands for breast cancer cells using the one-bead one-compound combinatorial method. *J. Med. Chem.* **2009**, *52*, 126−133.

(10) Hintersteiner, M.; Kimmerlin, T.; Kalthoff, F.; Stoeckli, M.; Garavel, G.; Seifert, J.-M. M.; Meisner, N.-C. C.; Uhl, V.; Buehler, C.; Weidemann, T.; Auer, M. Single bead labeling method for combining confocal fluorescence on-bead screening and solution validation of tagged one-bead one-compound libraries. *Chem. Biol.* **2009**, *16*, 724−735.

(11) Bailey, T. L. *Discovering motifs in DNA and protein sequences: The approximate common substring problem*. Ph.D. thesis, University of California at San Diego, 1995

(12) Bailey, T. L.; Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings International Conference on Intelligent Systems for Molecular Biology, ISMB*, ICSB: La Jolla, California, USA, Aug 14−17, 1994; Vol. 2, pp 28−36.

(13) Wu, J.; Aluko, R. E. Quantitative structure-activity relationship study of bitter di- and tri-peptides including relationship with angiotensin I-converting enzyme inhibitory activity. *J. Peptide Sci.* **2007**, *13*, 63−69.

(14) Hawkins, D. M.; Basak, S. C.; Mills, D. Assessing Model Fit by Cross-Validation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579−586.

(15) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.

(16) Macqueen, J. Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistical Probability*, Berkeley, CA, June 21−July 18 and Dec 27−Jan 7, 1965−1966; Regents of the University of California, 1967; pp 281−297.

(17) Goujon, M.; McWilliam, H.; Li, W.; Valentin, F.; Squizzato, S.; Paern, J.; Lopez, R. A new bioinformatics analysis tools framework at EMBL−EBI. *Nucleic Acids Res.* **2010**, *38*, W695−W699.

(18) Henikoff, S.; Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 10915−10919.

(19) Hartigan, J. A.; Wong, M. A. Algorithm AS 136: A K-Means Clustering Algorithm. *Appl. Stat.* **1979**, *28*, 100−108.

(20) R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing, Vienna, Austria, 2011.

(21) Misawa, K.; Tajima, F. A Simple Method for Classifying Genes and a Bootstrap Test for Classifications. *Mol. Biol. Evol.* **2000**, *17*, 1879−1884.

(22) Gordon, A. D. *Classification*, 2nd ed.; Chapman & Hall/CRC Monographs on Statistics & Applied Probability; Chapman and Hall/CRC: Boca Raton, FL, 1999.

(23) Bailey, T. L.; Elkan, C. Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization. *Mach. Learn.* **1995**, *21*, 51−80.

(24) Dempster, A. P.; Laird, N. M.; Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. Ser. B* **1977**, 1−38.

(25) Chen, X.; Ren, L.; Kim, S.; Carpino, N.; Daniel, J. L.; Kunapuli, S. P.; Tsygankov, A. Y.; Pei, D. Determination of the substrate specificity of protein-tyrosine phosphatase TULA-2 and identification of Syk as a TULA-2 substrate. *J. Biol. Chem.* **2010**, *285*, 31268−31276.

(26) Doweyko, A. QSAR: dead or alive? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 81−89.

(27) Bauer, D. F. Constructing Confidence Sets Using Rank Statistics. *J. Am. Stat. Assoc.* **1972**, *67*, 687−690.

(28) Ghose, A. K.; Crippen, G. M. Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 21−35.

(29) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493−500.