Identifying Affinity Classes of Inorganic Materials Binding Sequences via a Graph-based Model

Nan Du, Marc R. Knecht, Mark T. Swihart, Zhenghua Tang, Tiffany R. Walsh and Aidong Zhang

Abstract—Rapid advances in bionanotechnology have recently generated growing interest in identifying peptides that bind to inorganic materials and classifying them based on their inorganic material affinities. However, there are some distinct characteristics of inorganic materials binding sequence data that limit the performance of many widely-used classification methods when applied to this problem. In this paper, we propose a novel framework to predict the affinity classes of peptide sequences with respect to an associated inorganic material. We first generate a large set of simulated peptide sequences based on an amino acid transition matrix tailored for the specific inorganic material. Then the probability of test sequences belonging to a specific affinity class is calculated by minimizing an objective function. In addition, the objective function is minimized through iterative propagation of probability estimates among sequences and sequence clusters. Results of computational experiments on two real inorganic material binding sequence datasets show that the proposed framework is highly effective for identifying the affinity classes of inorganic material binding sequences. Moreover, the experiments on the SCOP (structural classification of proteins) dataset shows that the proposed framework is general and can be applied to traditional protein sequences.

Index Terms—Inorganic material, peptide sequences, classification

1 INTRODUCTION

Over the past decade, many studies have been published for analyzing the peptide sequences with affinity to biological entities such as enzymes, cells, viruses, lipids and proteins. Recently, interest in identifying and classifying peptides that interact specifically with inorganic materials has grown. These inorganic materials binding peptide sequences have been identified from biocombinatorial peptide libraries using phage display [1], cell surface display [2], and yeast display [3].

In particular, numerous studies have been reported about the peptide sequences that bind to the inorganic materials, such as noble metals (gold, silver, platinum) [4], [5], [6], [7], [8], semiconductors (zinc sulfide, cadmium sulfide) [9], [10], [11], [12], and metal oxides (silica, titanium and magnetite) [13], [14], [15], [16], [17], [18], [19], which are of great interest for applications in technology and medicine. Inorganic material binding peptide sequences, which are usually 7-14 amino acids long, are differentiated from other polypeptides by their specific molecular recognition properties for targeted inorganic material surfaces [20]. Effectively identifying the affinity classes, which shows the binding strength of a specific sequence with respect to the target inorganic material, is crucial for further designing novel peptides [21]. The binding affinity of a peptide to an inorganic surface is the result of a complex interplay between the binding strength of its individual residues and its conformation. The binding strength of a sequence for a specific material is usually measured with the adsorption free energy (∆Gads), which is then used to classify the affinity class as weak, medium, or strong for each sequence.

Despite extensive recent reports on combinatorially selected inorganic binding peptides and their bionanotechnology utility as synthesizers and molecular linkers [22], [23], [20], there is still limited knowledge about the relationships between binding peptide sequences and their associated inorganic materials. Therefore, by using machine learning technology to suggest sequence affinity classes, we can predict new sequences having desired affinity for specific inorganic materials, without doing new large-scale screenings via phage display.

Various approaches have been used or developed for recognizing both close and distant homologs of given protein sequences, which is one of the central themes in bioinformatics. Most of the work is based on established machine learning models such as Hidden Markov model (HMM) [24], [25], Neural Network (NN) [26], [27] and Support vector machine (SVM) [28]. However, the problem of inorganic material binding peptide sequence affinity classes identification has some distinct challenges that are rarely faced in protein sequence identification, which markedly limit the performance of the models mentioned above, despite their success in other types of protein sequences detection.

- Nan Du and Aidong Zhang are with the Computer Science and Engineering Department, University at Buffalo (SUNY), Buffalo, NY 14260. E-mail: nandu,azhang@buffalo.edu
- Marc R. Knecht and Zhenghua Tang are with Department of Chemistry, University of Miami, 1301 Memorial Drive, Coral Gables, Florida 33146. E-mail: knecht,z.tang@miami.edu
- Mark T. Swihart is with Department of Chemical and Biological Engineering, University at Buffalo (SUNY), Buffalo, NY 14260. E-mail: swihart@buffalo.edu
- Tiffany R. Walsh is with Institute for Frontier Materials, Deakin University, Geelong, Vic. 3216, Australia. E-mail: tiffany.walsh@deakin.edu.au
Challenge I: The number of labeled samples is usually insufficient. As an emerging topic, the peptide sequences identified for binding solid inorganic materials have been developed only in the last decade, and are not so well studied compared to protein sequences analysis which has much longer history. For example, unlike protein sequences analysis that has numerous large-scale public datasets such as GPCR [29] or SCOP [30], no complete result of large-scale screening experiments has been made publicly available for the inorganic material binding sequences. Therefore, unlike protein sequence research which has many public large databases and publicly available experiment results, the data about inorganic binding peptide sequences are usually quite few. Most existing protein sequence classification approaches require a large set of labeled samples to train an accurate model. However, labeling the affinity classes for a large number of inorganic material binding sequences is very time-consuming and expensive. Thus it is usually infeasible. If only a limited number of labeled samples are available for the model training, the learned model may suffer from the problems of over-fitting or under-fitting. As a machine learning method which has received much attention in the past decade, Semi-Supervised Learning (SSL) [31] is good at handling the lack of sufficient labeled training data problem. However, the utility of this method may be markedly limited due to the next challenge.

Challenge II: The peptide sequences belonging to the same affinity class may be very dissimilar. Usually, the protein sequences which belong to the same family follow some apparent patterns, in other words, they are similar to each other by some views. However, the “similarity” between inorganic material binding peptide sequences from the same affinity class may be not so apparent. In some cases, the intra-similarity which measures the similarity of all sequences inside the same class is even less than the inter-similarity which measures the similarity among the sequences from different classes. This phenomenon also means some peptide sequences belonging to the same class may be dissimilar with each other, at least by the current knowledge. This observation reflects the fact that the inorganic material binding sequences do not satisfy the smoothness assumption at the class level which is generally assumed in both supervised learning and semi-supervised learning.

In light of these challenges for inorganic material binding sequence affinity classes identification, we propose a novel framework which includes two parts. First, to tackle the insufficient data challenge, we augment the training sequence set with simulated sequences which are generated based on a new amino acid transition matrix. By using the simulated sequences, we incorporate not only the prior phylogenetic knowledge but also the specific sequence patterns responsible for the target inorganic material into the training data. Second, instead of searching the patterns globally from the peptide sequences belonging to the same affinity class, we separate the sequences into smaller clusters and try to learn the patterns from them locally via a graph-based optimization model. Intuitively, since there are few obvious patterns that could be found at the class level, we search for them at the smaller cluster level.

Based on the two strategies mentioned above, we propose a novel model that combines the sequence simulation and cluster-based sequence affinity identification. The initial idea was published in [32]. This paper extends the original idea to formulate a solid method and provide more supportive, comprehensive experiments. The main process of the proposed method is shown in Fig. 1, where we first use the labeled sequences as seeds to simulate more sequences, and then all the labeled and simulated sequences are used to train our graph-based optimization model which is effective at identifying the sequences’ affinity classes. We will discuss the proposed method in detail in the following sections.
2 RELATED WORK

As an emerging research topic, there is very little published work on identifying the affinity classes of inorganic material binding sequences that we can compare to. But as a similar topic, much research has been devoted to the question of identifying the homologs of the protein sequences. HMM is a widely-used probability modeling method for protein homology detection [24], [25], [33] which first generates a probability for each specific sequence family and then calculates the likelihood of an unknown sequence fitting each family. Another type of direct modeling methods for protein homology detection is based on Neural Network [26], [27], where the multilayer nature of neural network allows them to discover non-linear higher order correlations among the sequences. As a widely-used machine learning algorithm, SVM [28] has been also applied to protein homology detection problems. Mak et al. [34] proposed a SVM based model named PairProSVM to automatically predict the sub-cellular locations of proteins sequences. Karchin et al. [29] combined the HMM with the SVM to identify the protein homologies. Tian et al. proposed a weighted version of SVM to weaken the influence of outliers for improving protein sub-cellular localization predictions [35]. However, these methods are inappropriate in our case for two reasons. First, they ask for a training set consisting of sufficient labeled examples. Second, they try to learn the pattern from each class which may not exist at this level.

Moreover, besides the differences with the traditional classification approaches, the proposed framework is also different from the following work: 1) Oren et al. [21] has proposed a method to generate a new transition matrix and make the classification based on it. The first difference between the work presented here and Oren's work is that they only consider the sequence classification problem via learning the patterns from the entire sequence set belonging to the same affinity class. Second, the newly generated transition matrix in [21] was only used to calculate the pairwise distance between sequences. In our proposed method, the newly generated matrix is also used to generate the simulated sequences. 2) Ge et al. [36] proposed a consensus maximization model to solve the problem of finding informative genes from multiple studies. Although the proposed method has the same intuition as Ge’s work in which a cluster should correspond to a particular class if the majority of instances in this cluster belongs to class, it aimed at making the reliable prediction by utilizing multiple experimental results which is much different from our work. In our case, we only have the raw dataset (i.e. labeled inorganic material sequences) rather than multiple experimental results.

3 DATASETS AND PROBLEM DEFINITION

In this section, we describe the datasets used in this paper and present the problem definition.

3.1 Datasets

We have used three datasets to demonstrate the proposed method’s performance. The first dataset is from Oren et al. [21]. This dataset consists of a total of 25 quartz (rhombohedral silica, SiO2) binding peptide sequences which were identified using phage-display techniques. All these peptide sequences are further classified into two classes based on their affinity strength: strong and weak binder classes which contain 10 and 15 sequences, respectively. To better demonstrate the problem and show the proposed method in the rest of the paper, we abstract a sample set which includes two affinity classes from this dataset and show it in Table 1.

<table>
<thead>
<tr>
<th>Strong Class</th>
<th>Name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS202</td>
<td></td>
<td>RLNPFSQMDPPE</td>
</tr>
<tr>
<td>DS189</td>
<td></td>
<td>QTWWPPPLWSTS</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Weak Class</th>
<th>Name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS201</td>
<td></td>
<td>MEGQYKSNLLFT</td>
</tr>
<tr>
<td>DS191</td>
<td></td>
<td>VAPVQNLHFGA</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>

The second inorganic material binding peptide sequence dataset is from our systematic study of peptide binding on gold (Au) [37], combined with the previous data from Wang et al. [38], to give a total of 32 peptide sequences. Sequences in our sequence set following the pattern XHXXHXX, where X is an arbitrary amino acid are from Wang et al. [38]. Since any peptide sequences that containing cysteine (i.e. amino acid C) can bind strongly onto the gold surface, without loss of generality, any sequences contains cysteine are not considered. Using measured adsorption free energies (ΔG kJ/mol) for all the sequences, we drew the boundary between strong and weak binding sequences, such that the weak class has ΔG > −25 kJ/mol, and the strong class has ΔG ≤ −25 kJ/mol. Note that, Hnilova et al. [39] have shown that sequence ’TLRRWDRRNL’ (AUBP30) has weak binding ability to gold. Although they did not report the free energy for it, it is very likely to reside in the weak set based on the qualitative binding analysis. All the sequences from the strong and weak classes are listed in Table 2.

It is worth noticing that these datasets illustrate well the two challenges mentioned above. First, there are only around ten sequences available for each affinity class, which is very few in comparison to the data size used for classifier training in protein sequence analysis where hundreds or thousands of sequences are usually involved [33], [25]. Second, the unobvious pattern challenge shown in these datasets is illustrated well in Fig. 2 and Fig. 3. In this figure, based on the total similarity scores (TSS) defined in [21], we first calculate the total similarity of sequences from the same class A via the following equation:

\[ TSS_A = \frac{1}{NA \times (NA - 1)} \sum_{i=1}^{NA} \sum_{j=1}^{NA} PSS_{ij}(1 - \delta_{ij}), \]
where $\delta$ is the usual Kronecker delta function in which $\delta_{ij} = 1$ when $i = j$ and 0 otherwise, $NA$ is the total number of sequences in set $A$, and $PSS_{ij}$ is the similarity between the $i$th sequence and $j$th sequence of set $A$ calculated via the Needleman-Wunsch algorithm [40]. For the sake of simplicity, we call it self-class similarity for short. Moreover, the TSS of the sequences across the classes $A$ and $B$ are calculated as:

$$TSS_{A-B} = \frac{1}{NA \times NB} \sum_{i=1}^{NA} \sum_{j=1}^{NB} PSS_{ij},$$

where $NB$ is the total number of sequences in set $B$. Correspondingly, the total similarity for sequences across the classes is named across-class similarity for short.

To calculate $PSS_{ij}$, we need to provide a transition matrix on which the optimal scoring alignment would be made. Without loss of generality, we have used both the Pam 250 [41] (Fig. 2(a) and Fig. 3(a)) and Blosum 62 [42] (Fig. 2(b) and Fig. 3(b)) as the transition matrices, respectively. Fig. 2 shows that the sequences belonging to the weak class have very low or no significant similarities. Their self-similarity is much lower than the cross-class similarity. Similarly, as shown in Fig. 3, the similarities of the sequences belonging to the strong gold binding set are very close to the cross-class similarity. Due to this phenomenon, the traditional classification approaches cannot readily identify an effective pattern.

To demonstrate the proposed work is a general framework which is also effective on predicting the homology families of the traditional protein sequence, the third dataset: Structural Classification of Proteins SCOP dataset from [43] is also used. In addition, we employ the approach developed by Anoop Kumar and Lenore Cowen [25] to pick the SCOP families, where acquired proteins are further grouped into seven families (i.e. A, B, C, D, E, F and G). The size and the length of longest/shortest of amino acids at each family in the dataset are shown in Table 3, and the data we used are available at http://www.acsu.buffalo.edu/~nandu/InorganicSeq/.

Table 3: Summary of the protein sequence data

<table>
<thead>
<tr>
<th>Family</th>
<th>Number of Seq</th>
<th>Length of Shortest Seq</th>
<th>Length of Longest Seq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class A</td>
<td>23</td>
<td>160</td>
<td>177</td>
</tr>
<tr>
<td>Class B</td>
<td>23</td>
<td>72</td>
<td>136</td>
</tr>
<tr>
<td>Class C</td>
<td>16</td>
<td>224</td>
<td>260</td>
</tr>
<tr>
<td>Class D</td>
<td>19</td>
<td>120</td>
<td>144</td>
</tr>
<tr>
<td>Class E</td>
<td>18</td>
<td>324</td>
<td>429</td>
</tr>
<tr>
<td>Class F</td>
<td>14</td>
<td>131</td>
<td>221</td>
</tr>
<tr>
<td>Class G</td>
<td>20</td>
<td>45</td>
<td>85</td>
</tr>
</tbody>
</table>

3.2 Problem Definition

We consider our problem as identifying the affinity classes for the test inorganic material binding peptide sequences based on the training sequences. We start from a pool of $l+u$ peptide sequences $S_{source} = \{s_1, ..., s_l, ..., s_{l+u}\}$ where each sequence is represented as a series of ordered amino acids. To better understand what a peptide sequence looks like, let us take a peptide sequence from Table 1 as an example. D-S202: RLNPPSQMDPPF is a peptide sequence composed of twelve ordered amino acids where each letter denotes one of the 20 standard amino acids. We also assume in this sequence pool $S_{source}$, the first $l$ sequences are labeled $s_i \in S_{source}(1 \leq i \leq l)$ based on its affinity to the
target inorganic material (e.g. weak or strong), which together is named $L$, and the rest of $u$ sequences are unlabeled $s_i \in S_{source}(l + 1 \leq i \leq l + u)$ and together is named $U$, where $L \cup U = S_{source}$. Our goal is to predict the labels of peptide sequences in $U$, using the training sequences in $L$.

4 Peptide Sequence Simulation

As we mentioned above, lack of labeled data is a general problem we usually face when working on inorganic material binding sequences. One of the most successful methods to date for recognizing protein sequences based on evolutionary knowledge is using simulated sequences. Nowadays, there are many studies [25, 44] which have shown that augmenting the training set with the simulated sequences generated from an amino acids transition matrix such as Blosum 62 and Pam 250 can increase the homologs identification performance. One can reasonably expect that a set of peptides generated by directed evolution to recognize a given solid material will have similar sequences [21].

Although these transition matrices are shown to be efficient and gain wide acceptance, we cannot directly apply this technique to generate simulated sequences. These transition matrices are derived from the large-scale natural protein sequence databases rather than the target inorganic material binding sequences, which means these existing matrices could not represent the target inorganic material well. Thus, we only use a traditional transition matrix as a seed and based on it we generate a new transition matrix which not only maintains the prior knowledge from proteins but also captures the significant knowledge inside the target inorganic material.

Here, aiming to provide a more comprehensive and diverse view for our model, we use a two-step simulated sequence generation approach to enlarge our training set. First, we generate a new transition matrix which includes the source sequence graph (Section 5.1) where the relationships among sequences are better measured and many efficient clustering methods are available. Instead of searching for the patterns at the class level, we propose a graph-based optimization model to estimate the conditional probability of the test sequences belonging to each affinity class. Our method begins by mapping sequences from the sequence pool into nodes of a sequence-to-sequence graph (Section 5.1) where the relationships among sequences are better measured and many efficient clustering methods are available. Instead of searching for the patterns at the class level, we partition the sequences into clusters where we believe the significant patterns exist and an objective function (Section 5.2) is proposed to learn the conditional probability of each sequence belonging to a specific affinity class. Finally, we present

$$P_{ij} = \frac{M'_{ij}}{\sum_{j=1}^{20} M_{ij}}.$$

Note that, all the probabilities are calculated after normalization of the values in $M^*$ into a positive value space (e.g. 0–1). As an example, Fig. 4 shows the process of mutating an amino acid in the selected position based on the mutation probability, and we keep replacing the amino acids in the target sequence until a desired mutation threshold $t$ is reached [25].

![Fig. 4: An example of mutating an amino acid in the selected position. When a specific position (e.g. 8-th) is selected from the target sequence, the corresponding amino acid M is mutated to be another amino acid W based on the mutation probability.](image)

By this two-step method, we can incorporate not only the prior phylogenetic knowledge but also the specific amino acid pattern responsible for binding to the target inorganic material into the data. Accordingly, based on this peptide sequence simulation method, for each labeled source sequence $s_i \in S_{source}(1 \leq i \leq l)$, we generate $m$ mutated sequences, which is represented as a simulated peptide sequence set $S_{simulated} = \{s'_1, ..., s'_{n\times m}\}$. Finally, we define the sequence pool as $S = S_{source} \cup S_{simulated}$ which includes the source peptide sequences and simulated sequences. We will show that the simulated sequences effectively improve the performance in the experiments.

5 Graph-based Optimization Model

Aiming to handle the challenge that the obvious patterns are hard to find at the class level, we propose a graph-based optimization model to estimate the conditional probability of the test sequences belonging to each affinity class. Our method begins by mapping sequences from the sequence pool into nodes of a sequence-to-sequence graph (Section 5.1) where the relationships among sequences are better measured and many efficient clustering methods are available. Instead of searching for the patterns at the class level, we partition the sequences into clusters where we believe the significant patterns exist and an objective function (Section 5.2) is proposed to learn the conditional probability of each sequence belonging to a specific affinity class. Finally, we present
an efficient iterative algorithm to obtain the optimal value of the objective function (Section 5.3).

5.1 Mapping Sequences into Nodes of a Graph

We map all the sequences into a graph where each node denotes a peptide sequence and each edge denotes the pairwise similarity between two sequences. This graph offers a good understanding of the pairwise relationships among peptide sequences and is easily partitioned into clusters. The pairwise similarity among sequences is calculated using Needleman-Wunsch [40] algorithm after local alignment between each sequence pair using Smith-Waterman algorithm [45].

5.2 The Objective Function

The key idea of our approach is that, instead of searching for patterns at the class level, we narrow down the affinity class prediction problem from the class level to clustering problems, to partition the sequence-to-sequence similarity matrix that, each entry in the belongingness matrix corresponds to patterns at the class level, we narrow down the knowledge of clusters.

Belongingness Matrix: We denote the belongingness matrix \( \mathbf{D} \) of \( ij \)-th entry in the matrix \( \mathbf{D} \), and \( d_{ij} \) denote vectors of \( i \)-th row and \( j \)-th column of matrix \( \mathbf{D} \), respectively.

Sequence Probability Matrix: The conditional probability of peptide sequence \( s_i \) belonging to class \( z \) \( (s_{iz} = P(y = z|s_i)) \) is estimated with an \( N \times D \) matrix \( \mathbf{S} \) where \( D \) is the number of affinity classes we want to classify.

Cluster Probability Matrix: The conditional probability of cluster \( c_j \) belonging to class \( z \) \( (c_{jz} = P(y = z|c_j)) \) is estimated as a \( V \times D \) matrix \( \mathbf{C} \), where \( c_{jz} \) represents the probability of a cluster \( c_j \) belonging to a class \( z \).

Sequence Labeled Matrix: In the labeled sequence set \( L \) the sequences have the initial class labels which are represented by an \( N \times D \) matrix \( \mathbf{F} \), where \( f_{iz} = 1 \) if we know sequence \( s_i \) belonging to class \( z \) in advance, and 0 otherwise.

Cluster Labeled Matrix: We may also have prior information of a cluster belonging to a specific class. We use a \( V \times D \) matrix \( \mathbf{Y} \) to define initial labels for clusters where \( y_{jz} = 1 \) denotes that we are confident that a cluster \( c_j \) belongs to a specific class \( z \), and 0 otherwise.

Specifically, we assign a cluster \( c_j \) to a specific class \( z \) if all the source sequences in it belong to the same class \( z \).

Cluster Similarity Matrix: In addition, a \( V \times V \) matrix \( \mathbf{W} \) denotes the similarity among the sequence clusters, where \( w_{ij} \) is the similarity between the sequence cluster \( c_i \) and \( c_j \). Specifically, the pairwise cluster similarity is calculated using \( TSS_{A-B} \) between the sets of sequence binders \( A \) and \( B \).

Now we formulate the affinity class identification problem as the following objective function:

\[
\min_{S,C} J(S,C) = \min_{S,C} \left( \sum_{i=1}^{N} \sum_{j=1}^{V} b_{ij} \| \mathbf{s}_{iz} - \mathbf{c}_{jz} \|^2 + \alpha \sum_{i=1}^{V} \sum_{j=1}^{V} w_{ij} \| \mathbf{c}_{iz} - \mathbf{c}_{jz} \|^2 + \beta \sum_{i=1}^{N} h_{i} \| \mathbf{s}_{iz} - \mathbf{f}_{iz} \|^2 + \gamma \sum_{j=1}^{V} k_{j} \| \mathbf{c}_{jz} - \mathbf{y}_{jz} \|^2 \right)
\]

subject to the following conditions:

\[
\sum_{z=1}^{D} s_{iz} = 1, \quad s_{iz} \geq 0 \quad \sum_{z=1}^{D} c_{jz} = 1, \quad c_{jz} \geq 0,
\]

where \( \| \cdot \|^2 \) indicates the \( L_2 \) norm. The first term in Eq. (4), \( \sum_{i=1}^{N} \sum_{j=1}^{V} b_{ij} \| \mathbf{s}_{iz} - \mathbf{c}_{jz} \|^2 \), ensures that a sequence should have similar probability vector as the cluster it belongs to, namely, cluster \( c_j \) should correspond to class \( z \) if the majority of sequences in this cluster belong to class \( z \). Intuitively, the higher the deviation, the larger penalty would get. The second term \( \alpha \sum_{i=1}^{V} \sum_{j=1}^{V} w_{ij} \| \mathbf{c}_{iz} - \mathbf{c}_{jz} \|^2 \) corresponds to the intuition that the clusters which are close to each other should have similar class, and \( \alpha \) denotes the confidence over this source of information. From the view of graph theory, this term is propagating the class information among the clusters. The third term \( \beta \sum_{i=1}^{N} h_{i} \| \mathbf{s}_{iz} - \mathbf{f}_{iz} \|^2 \) applies the constraint that the predictions should not deviate too much from the corresponding sequence ground-truth and \( \beta \) is the parameter that expresses the confidence of our belief on the prior knowledge of sequences. Similarly, the last term \( \gamma \sum_{j=1}^{V} k_{j} \| \mathbf{c}_{jz} - \mathbf{y}_{jz} \|^2 \) is the loss function penalizing the deviation between predictions and our prior knowledge of clusters, and \( \gamma \) is the parameter that expresses the confidence of our belief on the prior knowledge of clusters.

5.3 Iterative Update Algorithm

It is easy to prove that the objective function Eq. (4) is convex which makes it possible to find a global optimal solution. To obtain the optimal solution for matrices \( S \) and \( C \), we propose to solve Eq. (4) using the block coordinate descent method [47]. At iteration \( t \), fixing the value of \( s_{iz}^t \), we can take the partial derivative to \( c_{jz}^t \) in Eq. (4) and set it to 0, and then obtain the update Formula Eq. (5):
\[ c_{ij}^T = \frac{\sum_{i=1}^{n} b_{ij} s_{ij}^{t-1} + \gamma k_{ij}}{\sum_{i=1}^{n} b_{ij} + \gamma k_{ij} + \alpha (t_j - l_j)}. \]  

Accordingly, the update can be represented as a matrix form as Eq. (6), where \( D_v = \text{diag} \{(\sum_{i=1}^{N} b_{ij})\} \) is the normalization factor, \( K_v = \text{diag} \{(\sum_{j=1}^{D} y_{ij})\} \) indicates the constraints for the clusters and \( \text{diag} \) denotes the diagonal elements of a matrix. Furthermore, \( \tilde{L} \) is the normalized laplacian [48] defined as \( \tilde{L} = D_w^{-\frac{1}{2}} W D_w^{-\frac{1}{2}} \), where \( D_w \) is the diagonal degree matrix of \( W \).

\[ C^t = (D_v + \gamma K_v + \alpha (I - \tilde{L}))^{-1}(A^T S^{t-1} + \gamma K_v Y). \]  

The Hessian matrix with respect to \( C \) is a diagonal matrix with entries \( \sum_{i=1}^{n} b_{ij} + \alpha > 0 \) and \( I - \tilde{L} \). The diagonal matrix is positive definite and it is easy to prove that \( I - \tilde{L} \) is also a semi-positive definite. Thus, the hessian matrix is a positive definite matrix, which means derivative for \( C \) gives the unique minimum of Eq. (4). Similarly, we can obtain the update formula Eq. (7) with respect to \( s_i \) through fixing \( c_{ij}^T \).

\[ s_i^T = \sum_{b_{ij}} b_{ij} c_{ij}^T + \beta h_i f_i. \]  

Also, the matrix form of Eq. (7) is as following:

\[ S^t = (D_n + \beta H_n)^{-1}(A C^t + \beta H_n F), \]  

where \( D_n = \text{diag} \{(\sum_{i=1}^{N} b_{ij})\} \) is the normalization factor and \( H_n = \text{diag} \{(\sum_{j=1}^{D} f_{ij})\} \) indicates the constraints for the sequences. The hessian matrix is also a diagonal matrix with diagonal elements \( \sum_{j=1}^{n} b_{ij} > 0 \), which means the derivative of \( S \) gives the unique minimum of Eq. (4). To sum up, the pseudo-code of iteratively solving Eq. (4) by the block coordinate descend method is shown as Algorithm 1, where \( \epsilon \) is a convergence threshold. Because the proposed method is based on a graph model, we name our approach Peptide Sequences Identification Graph Model - PSIGM.

![Algorithm 1](image)

5.4 Time Complexity
The time complexity of the proposed algorithm is composed of two parts: updating the cluster probability matrix \( C \) and updating the sequence probability matrix \( S \). For updating the matrix \( C \), the time complexity is \( O(V N^2 D + V^3 + V^2 D) \) where \( N \) is the size of the peptide sequence pool and \( V \) is the number of clusters and \( D \) is the number of affinity classes. Because in our case the sequence set is usually much larger than the number of clusters, thus the time complexity for the first step is \( O(V N^2 D) \). For updating the matrix \( S \), the time complexity is \( O(N^2 D + N V D) \), thus \( O(N^2 D) \). Therefore, the overall time complexity is \( O(V N^2 D) \). Suppose the number of iterations is \( k \), the time complexity of whole algorithm is \( O(k V N^2 D) \). In experiments, we observe that \( k \) is usually between 8 and 20.

6 EXPERIMENTS
In the following, we first conduct the experiments on both the quartz and gold binding sequence datasets to show that PSIGM is effective for identifying the binding affinity classes of inorganic material binding sequences, and then the experiments on the SCOP protein sequence dataset to show that PSIGM is a general framework which also works effectively in other kinds of sequence sets. Because most of our baselines are designed as binary classifiers, for the sake of simplicity, in the following experiments we only consider the case of weak and strong binder identification of the datasets mentioned in Section 3.1, although our proposed method is not restricted to binary classification. Throughout all the experiments, we set \( \alpha = 2 \), \( \beta = 10 \), and \( \gamma = 2 \) as default values of our algorithm. The rationale is that both \( \alpha \) and \( \gamma \) depend on the clustering result which is influenced by some uncertain factors such as the number of clusters and the initial center of each cluster, thus assigning a relative low value to them is better; on the other hand, \( \beta \) shows our confidence in the labeled sequences which come from strict and reliable experiments, thus \( \beta \) should be assigned a relative larger value.
Fig. 5: An example of illustrating the label propagation at each iteration. (A) partition of all the nodes (each sequence is represented as a node here) into multiple clusters; (B) conditional probability estimate of clusters (i.e. cluster probability matrix $C$) receiving the label information from the sequences (nodes) belonging to them; (C) each cluster propagates its class information to their neighboring clusters; and (D) after updating probability, each cluster passes the label information back to its members conditional probability (i.e. sequence probability matrix $S$).

6.1 Experiments on Material Binding Sequences

To show our proposed framework is effective on predicting the binding affinity class of inorganic material binding sequences, we perform the following three experiments to demonstrate that: 1) simulated sequences and the information propagation among the clusters can effectively alleviate the limitation due to challenge I; 2) searching the patterns from clusters rather than classes helps in handling challenge II; and 3) the newly generated transition matrix contributes to our proposed method’s performance. For each accuracy (i.e. the percent of testing set examples correctly classified by the classifier when compared with the ground truth) shown in the following experiments, we performed experiments 5 times using Leave-one-out validation and report the mean value. We iteratively select one labeled sequence as the test sequence, and use the rest of the labeled sequences to generate the simulated sequences and train the model. When the model is well trained, we can predict the test sequences affinity class. The reason why we use the Leave-one-out validation rather than cross validation in the inorganic material binding sequences is that the number of labeled sequences is insufficient. In such a case, each one of them may represent significant underlying pattern or characteristic. In addition, for each experiment, we fix the threshold $\epsilon$ for convergence to $10^{-4}$.

The effect of simulated sequences and cluster information propagation. To test the effect of simulated sequences and information propagation among the clusters, we ran our method with or without using simulated sequences or information propagation among clusters, respectively. Furthermore, as we mentioned in Section 5.2, we need to partition all the sequences into $V$ clusters, thus we also want to show the relationship between the proposed method’s performance and the number of clusters. We vary the number of clusters as 2, 5, 10, 15, 18 and 20.

We show that both strategies, sequence simulation and information propagation among the clusters, are crucial in improving the performance. The result is shown in Fig. 6, where the $x$ axis denotes the different number of clusters and $y$ axis denotes the accuracy. We can see that simulated sequences contribute to the performance improvement. Also, in the absence of the information propagation among clusters, the performance degrades. Finally, we notice that the hill-like shapes appear as the number of clusters increasing for all the three cases. Most likely, when the number of clusters is too low, it is close to a global view; when it is too high, the clusters would be too trivial to learn from.

Performance comparison with baselines. In this part, we compare the proposed method with 5 other algorithms mentioned above including SVM, Neural Network, HMM, Learning with local and global consistency (LLGC) [48], which is a well-known graph-based Semi-Supervised
Learning algorithm. For fairness and comprehensiveness, we have also tried adding the simulated sequences used for our framework to these methods which are marked as SVM*, Neural Network* and HMM*. Note that, since LLGC was designed as a semi-supervised algorithm which needs unlabeled instances to aid propagating the labeled information, thus we only consider the LLGC with the simulated sequences which are used as unlabeled data. Thus, all the methods are separated into two parts: using the simulated sequences and without using the simulated sequences. To measure the influence from the different mutation rate \( t \) which is used to generate simulated sequences, we vary the mutation rate \( t \) at 5%, 10%, 15% and 20%. The results of predicting the quartz binding affinity classes comparing with baselines on the two inorganic binding sequences dataset are shown in Table 4 and Table 5, respectively.

Note that the proposed method significantly outperforms the others in predicting the affinity classes of the test inorganic material binding sequences in most cases. In addition, the performance of the proposed method is not so sensitive to the settings of the mutation rate over the range considered. It is worth noticing that, instead of aiding the performance, the simulated sequences in SVM*, HMM* and NN* make the performance worse than without them. The main reason for this phenomenon is that: all these methods use a global view on the training data which is only represented among clusters. The clusters of the sequences are obtained from arbitrary clustering methods, which are not very stable. In other words, it may not be completely correct. Therefore, smaller \( \alpha \) usually yields better performance. \( \beta \) shows our confidence on the prior knowledge of the sequence classes. These sequence classes, which are obtained from serious physical or chemical experiments, are deemed to be reliable and thus a large \( \beta \) is usually better. \( \gamma \) denotes the confidence on the prior knowledge of cluster classes. This information may not be totally reliable, therefore lower value usually yields better results. The results in Fig. 7 confirm our observation.

**Parameter Sensitivity.** There are three parameters in our objective function Eq. 4: \( \alpha, \beta \) and \( \gamma \). We conducted sensitivity experiments, shown in Fig. 7. In the experiments, when one parameter is varied, the other two parameters are fixed at their default settings (i.e., \( \alpha = 2, \beta = 10, \) and \( \gamma = 2 \)). Note that, \( \alpha \) represents the confidence of our belief over information propagation among clusters. The clusters of the sequences are obtained from arbitrary clustering methods, which are not very stable. In other words, it may not be completely correct. Therefore, smaller \( \alpha \) usually yields better performance. \( \beta \) shows our confidence on the prior knowledge of the sequence classes. These sequence classes, which are obtained from serious physical or chemical experiments, are deemed to be reliable and thus a large \( \beta \) is usually better. \( \gamma \) denotes the confidence on the prior knowledge of cluster classes. This information may not be totally reliable, therefore lower value usually yields better results. The results in Fig. 7 confirm our observation.
mutation rate $t$ at 5%, 10%, 15% and 20%. We have compared the new transition matrix M1 which was generated based on Blosum 62 and M2 which was generated based on Pam 250 with four other widely-used transition matrices including Blosum 62, Pam 250, Dayhoff [49] and Gonnet [50] in Table 6. The result shows that, the newly generated transition matrices perform better than the others at each mutation rate.

### 6.2 Experiments on Protein Sequences

The proposed PSIGM is a general framework which is not limited to identifying the affinity class of inorganic material binding sequences. To prove that, we have used the SCOP protein data mentioned in Section 3.1. Instead of predicting the sequences’ affinity classes, we consider the problem in homology family prediction: for a specific family, could the proposed framework identify the sequences belonging to it from the remaining families? Correspondingly, we construct seven identification tasks from this dataset, where the sequences from one particular family are used as the positive set and the sequences from the remaining six families are used as the negative set. For example, when the sequences in family A are used as the positive set, the sequences from families B, C, D, E, F and G would be used as the negative set. Two experiments are performed to demonstrate that: 1) our PSIGM is a general framework which can also handle the tradition protein sequence identification; and 2) a moderate setting of mutation rate is conductive to improve the performance.

**Performance comparison with baselines.** It is worth noticing that, through handling the data in this way, it obtains the characteristics of inorganic binding sequences to some extent. Note that each result shown in the following experiments (i.e. Table 7 and Fig. 9) is the average of 10 times performance through 5-fold cross validation. Since the protein sequence dataset has relative sufficient training samples and the sequences that belong to the same protein family are similar to each other, we have used cross validation rather than Leave-one-out validation. Table 7 shows the result of predicting the homology family comparing with baselines which are mentioned in Section 6.1. As the table shows, the proposed method outperforms the other methods at each protein family’s prediction. Note that, the accuracies of predicting the homology families are much higher than the accuracies of predicting the affinity classes of the inorganic material binding sequences. The reasons behind this can be well explained by Fig. 8, which shows the self-class similarity of each prediction task. As we know, the more cross-class similarity surpasses the self-similarity, the more difficult two classes are separated.

### 7 Conclusion and Future Work

Identifying the affinity classes of peptide sequences binding to a specific inorganic material is a new and challenging research problem with broad applications. In this paper, we proposed a novel framework, PSIGM, to solve this problem. We begin with providing a two-step simulated peptide sequences generation method to make the training set more comprehensive and diverse. Moreover, unlike traditional machine learning approaches used for protein sequences identification that try...
to find the patterns from the class level, our framework partitions the sequences into smaller clusters and learns the patterns from them through using a graph-based optimization model. Extensive experimental studies demonstrate that the proposed framework can effectively identify the affinity classes of the inorganic material binding sequences.

In the future, to achieve better performance, we plan to use a cyclic model to validate and retrain PSIGM: first, we will select some sequences that have the most/least probabilities binding to a target inorganic material as a candidate set by using PSIGM; second, we plan to use some efficient experimental methods to validate the candidate sequence set such as QCM (Quartz Crystal Microbalance); finally, the validated sequences will be used to retrain the PSIGM, and then new candidate sequences will be selected from the sequence database based on their affinity, so on so forth. We believe that by this cyclic validation model, we can not only further validate PSIGM’s effectiveness but also keep retraining it to be better and better.

8 Acknowledgments

This material is based upon work supported by the Air Force Office of Scientific Research (AFOSR), grant number FA9550-12-1-0226. We gratefully acknowledge the Victorian Life Sciences Computation Facility (VLSCI) for allocation of computational resources. TRW thanks veski for an Innovation Fellowship.

Fig. 8: Total similarity scores of the self-class and the cross-class for each prediction task based on Pam 250. (A) self-class and cross-class TSS of class A and non-A; (B) self-class and cross-class TSS of class B and non-B; (C) self-class and cross-class TSS of class C and non-C; (D) self-class and cross-class TSS of class D and non-D; (E) self-class and cross-class TSS of class E and non-E; and (F) self-class and cross-class TSS of class F and non-F; (G) self-class and cross-class TSS of class G and non-G.

<table>
<thead>
<tr>
<th>Class</th>
<th>Self-Class</th>
<th>Cross-Class</th>
<th>Self-Class</th>
<th>Cross-Class</th>
<th>Self-Class</th>
<th>Cross-Class</th>
<th>Self-Class</th>
<th>Cross-Class</th>
<th>Self-Class</th>
<th>Cross-Class</th>
<th>Self-Class</th>
<th>Cross-Class</th>
<th>Self-Class</th>
<th>Cross-Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>[0, 100]</td>
<td>[0, 200]</td>
<td>[0, 300]</td>
<td>[0, 400]</td>
<td>[0, 500]</td>
<td>[0, 600]</td>
<td>[0, 700]</td>
<td>[0, 800]</td>
<td>[0, 900]</td>
<td>[0, 1000]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>[0, 200]</td>
<td>[0, 400]</td>
<td>[0, 600]</td>
<td>[0, 800]</td>
<td>[0, 1000]</td>
<td>[0, 1200]</td>
<td>[0, 1400]</td>
<td>[0, 1600]</td>
<td>[0, 1800]</td>
<td>[0, 2000]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>[0, 100]</td>
<td>[0, 200]</td>
<td>[0, 300]</td>
<td>[0, 400]</td>
<td>[0, 500]</td>
<td>[0, 600]</td>
<td>[0, 700]</td>
<td>[0, 800]</td>
<td>[0, 900]</td>
<td>[0, 1000]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>[0, 200]</td>
<td>[0, 400]</td>
<td>[0, 600]</td>
<td>[0, 800]</td>
<td>[0, 1000]</td>
<td>[0, 1200]</td>
<td>[0, 1400]</td>
<td>[0, 1600]</td>
<td>[0, 1800]</td>
<td>[0, 2000]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>[0, 100]</td>
<td>[0, 200]</td>
<td>[0, 300]</td>
<td>[0, 400]</td>
<td>[0, 500]</td>
<td>[0, 600]</td>
<td>[0, 700]</td>
<td>[0, 800]</td>
<td>[0, 900]</td>
<td>[0, 1000]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>[0, 200]</td>
<td>[0, 400]</td>
<td>[0, 600]</td>
<td>[0, 800]</td>
<td>[0, 1000]</td>
<td>[0, 1200]</td>
<td>[0, 1400]</td>
<td>[0, 1600]</td>
<td>[0, 1800]</td>
<td>[0, 2000]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>[0, 100]</td>
<td>[0, 200]</td>
<td>[0, 300]</td>
<td>[0, 400]</td>
<td>[0, 500]</td>
<td>[0, 600]</td>
<td>[0, 700]</td>
<td>[0, 800]</td>
<td>[0, 900]</td>
<td>[0, 1000]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

References

Nan Du received a B.S. degree from Guangdong University of Technology in 2006. After that, he received his M.S. degree from Southern China University of Technology in 2009. Since 2009, he has been working toward the Ph.D. degree in State University of New York at Buffalo, NY, with supervision by Prof. Aidong Zhang. His research interests are in the area of data mining, machine learning and bioinformatics.

Marc R. Knecht earned a B.S. degree in Chemistry from Duquesne University in 2001. In 2004, he received a Ph.D. in Bio-Inspired Chemistry from Vanderbilt University under the direction of Professor David W. Wright, followed by postdoctoral research at the University of Texas with Professor Richard M. Crooks focused on characterizing the structure/function relationship of nanocatalysts. After completing postdoctoral studies, he began his independent career as an assistant professor of Chemistry at University of Kentucky. In the summer of 2011, Professor Knecht joined the Department of Chemistry at the University of Miami as an associate professor. During his independent career, Professor Knecht has established a research program focused on elucidating the effects of the biotic/abiotic interface of bio-inspired nanomaterials. In this regard, his group has employed high-resolution characterization, activity studies, and synthetic analyses of peptides to demonstrate that the biological surface of bionanomaterials possesses significant control over the functionality and could serve as modification sites to control the activity. He has published 47 publications in this area.

Mark T. Swihart is a Professor in the Department of Chemical and Biological Engineering at the University at Buffalo (SUNY). He earned a B.S. in Chemical Engineering from Rice University in 1992, and a Ph.D. in Chemical Engineering in 1997 from the University of Minnesota. He then spent one year as a postdoctoral researcher in Mechanical Engineering at the University of Minnesota before joining the University at Buffalo as an assistant professor in 1998. Since 2007, he has directed a university-wide strategic initiative in Integrated Nanostructured Systems. His research interests include synthesis, processing, and applications of nanoparticles and other nanomaterials, and he has co-authored more than 120 journal papers in these areas. Dr. Swihart is a recipient of the Kenneth Whitby award from the American Association for Aerosol Research, the Schoellkopf medal from the Western New York section of the American Chemical Society, and the J.B. Wagner award from the Electrochemical Society.

Zhenghua Tang currently is a postdoctoral research associate working in Marc R. Knecht group at University of Miami. He obtained his B.S. degree at college of Chemistry and Chemical Engineering, Lanzhou University, Lanzhou, Gansu, P. R. China in 2005. He attended graduate school there from Aug. 2005 to Jun. 2007. During his graduate study, he went to Institute of Chemistry, Chinese Academy of Science (ICCAS) as a visiting student for about one year (2006-2007). In August 2007, he moved to US and obtained his Pd. D degree in chemistry from Department of Chemistry, Georgia State University in July, 2012. He started his current position since August, 2012. His research interest focuses on bio-inspired nanomaterials for targeted applications, including bio-nanocombinatorics, self-assembly, catalyst, multifunctional design and so on. He is the recipient of 2010 chairs award in Chemistry Department at GSU as well as 2011 Chinese Government Award for Outstanding Self-Financed Students Abroad.