Boone et al. BMC Bioinformatics (2018) 19:469
https://doi.org/10.1186/s12859-018-2514-6

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Antimicrobial peptide similarity and classification through rough set theory using physicochemical boundaries

Kyle Boone[1], Kyle Camarda[2], Paulette Spencer[3] and Candan Tamerler[4*]

Abstract

Background: Antimicrobial peptides attract considerable interest as novel agents to combat infections. Their long-time potency across bacteria, viruses and fungi as part of diverse innate immune systems offers a solution to overcome the rising concerns from antibiotic resistance. With the rapid increase of antimicrobial peptides reported in the databases, peptide selection becomes a challenge. We propose similarity analyses to describe key properties that distinguish between active and non-active peptide sequences building upon the physicochemical properties of antimicrobial peptides. We used an iterative supervised machine learning approach to classify active peptides from inactive peptides with low false discovery rates in a relatively short computational search time.

Results: By generating explicit boundaries, our method defines new categories of active and inactive peptides based on their physicochemical properties. Consequently, it describes physicochemical characteristics of similarity among active peptides and the physicochemical boundaries between active and inactive peptides in a single process. To build the similarity boundaries, we used the rough set theory approach; to our knowledge, this is the first time that this approach has been used to classify peptides. The modified rough set theory method limits the number of values describing a boundary to a user-defined limit. Our method is optimized for specificity over selectivity. Noting that false positives increase activity assays while false negatives only increase computational search time, our method provided a low false discovery rate. Published datasets were used to compare our rough set theory method to other published classification methods and based on this comparison, we achieved high selectivity and comparable sensitivity to currently available methods.

Conclusions: We developed rule sets that define physicochemical boundaries which allow us to directly classify the active sequences from inactive peptides. Existing classification methods are either sequence-order insensitive or length-dependent, whereas our method generates the rule sets that combine order-sensitive descriptors with length-independent descriptors. The method provides comparable or improved performance to currently available methods. Discovering the boundaries of physicochemical properties may lead to a new understanding of peptide similarity.

Keywords: Antibacterial peptides, Classification, Machine learning, Physicochemical properties, Rough set theory, Sequence similarity, Supervised learning, Functional peptide search

* Correspondence: ctamerler@ku.edu
[4]Mechanical Engineering Department, Bioengineering Program, Institute of Bioengineering Research, University of Kansas, Learned Hall, Room 3135A, 1530 W 15th St, Lawrence, KS 66045, USA
Full list of author information is available at the end of the article

Background

In the US, over 23,000 deaths each year are associated with drug-resistant bacterial infections [1]. These types of infections are central to the projected increase in deaths globally by 2050, which are expect to reach 10 million annually [2, 3]. The rise of antibiotic-resistant bacteria has prompted increasing interest in antimicrobial peptides as a solution to this critical issue [4]. Over 2800 antimicrobial peptides have been discovered from natural sources in the last decade [5–11]. Antibacterial peptides derived from these natural sequences have shown both broad-spectrum and improved activity against targeted bacteria [12–16]. Antibacterial peptide-mimics are introduced as another source to the existing peptide libraries by incorporating additional backbone chain atoms for more structural flexibility and resistance to protease degradation [17–20]. This list extends by exploring the post-translationally modified antimicrobial peptides offering chemical properties beyond the naturally occurring amino acids [21, 22].

While many antimicrobial peptides have been discovered at the laboratory bench, computational methods have been integrated into this search to find many more candidates. Encrypted antimicrobial peptides are an example in which known active peptides are queried against DNA repositories to find new antimicrobial peptides [23]. Among many methods, grammar-based methods and regular-expression-based match sequence patterns are used to identify functional similarity [24, 25]. Computer-aided molecular design [26–29] approaches using quantitative sequence activity relationships [30–33] (QSAR) predict the antibacterial level of peptides given key chemical properties. Artificial neural networks (ANN) have been used both to generate new sequences and to distinguish between active and inactive sequences [25, 34–37]. They are often used in the classification of antimicrobial peptide sequences [7, 38]. While ANNs are flexible enough to model many kinds of complex relationships, they lack transparency about how classification choices are made. Determining the boundaries of the similar antimicrobial peptide clusters remains difficult despite many existing machine learning methods.

Due to the ongoing need for improved antimicrobial peptide selection and design, many classification approaches have been developed with supervised machine learning methods. A recent review by Porto et al contrasts two different kinds of sequence representations for antibacterial classification [25]. The first kind of representation preserves the order of the sequence which tends to lead to length-dependent predictions [39]. False positives may be produced if the overall chemical properties of an antibacterial peptides are changed by adding amino acids with contradictory chemical properties. The second kind of sequence representation preserves overall sequence properties which tends to lead to order-insensitivity. False positives may be produced if the order of an active peptide is scrambled [24].

AntiBP [40] was one of the first online available services for antibacterial peptide prediction. AntiBP uses a sliding window of 15 residues to predict the classification using support vector machines (SVM) [41], quantitative matrices (QM) [42] and artificial neural networks (ANN) [43]. The strength of this approach is that the order of amino acids impacts the prediction. However, the weakness to having a constant window of amino acids is that the predictions are peptide-length dependent [39]. To overcome the peptide length dependence, another method CAMP (Collection of Antimicrobial Peptides) [44] was employed to use descriptors that summarize composition, physicochemical properties and structural features of the peptides. CAMP uses multiple machine learning approaches for these features such as SVM [45], ANN [46, 47], discriminate analysis (DA) [48] and random forest (RF) [49]. However, the descriptor approach is insensitive to the sequence order arrangement. For example, full-length sequence descriptors can be sensitive to the overall charge of a peptide but not its charge distribution. iAMP-2 L (antimicrobial peptide prediction two-level) [50] partially addresses the order insensitivity by calculating the autocorrelation of amino acid property values within the amino acid sequence. Other descriptors do not account for the order of the sequence [24]. Because the iAMP-2 L classification algorithm is based on a fuzzy *K*-nearest neighbor algorithm, clusters that are invariant for descriptors that include correlations would be sequence-order insensitive. This approach is also sequence-order insensitive to sequence rearrangements that preserve the correlation structure from the original peptide. Evolutionary Feature Construction [51–53] (EFC) method addresses this need by achieving order-sensitive classification by combining order sensitivity and length independence by selecting common chemical property sequence patterns for antimicrobial peptides. Length-independent classification is achieved with a support-vector machine method through physicochemical descriptors selected by FCBF (Fast-Correlation Based Filter selection) [52]. While this method does combine order-sensitivity and length-independence, it does not completely address either of these issues. Order-insensitivity is possible based on the rearrangements of amino acids that are indistinguishable by the pattern recognition scheme of compressing 20-amino acids into four categories.

We propose a novel method that addresses order sensitivity by calculating the physicochemical properties of sub-sequences in addition to using descriptors of physicochemical properties for length independence. Our method therefore combines order-sensitivity and length

independence as a new approach. We analyze these descriptors using rough set theory (RST). Rough set theory is a heuristic method for discovering rules, which distinguish between outcomes. These rules show which data features and data values are useful to distinguish between outcomes. To the best of our knowledge, RST has not yet been studied to classify peptide or protein sequences based on their activity. Our RST implementation uses features that summarize the physicochemical properties of the full-length sequences, which are sequence-order insensitive, and features which summarize constant-length subsequences, which are sequence-order sensitive. RST selects combinations of both kinds of descriptors into a single rule. Each rule defines its own cluster including the classification of the peptide's activity or inactivity.

Using a rough set theory approach that combines the algorithm of MLEM2 (modified learning from examples module, Version 2) [54] with the algorithm IRIM (Interesting Rule Induction Module) [55], we developed a method that investigates the sequence-function relationships. The main difference in from other RST methods is that it uses local coverings to generate rules, which are different from the lower and upper approximations in the basic RST methodology. IRIM is a method that optimizes for rules that have the most training set sequences that apply. This is different from MLEM2 in that IRIM may not provide a rule that applies to every training set sequence. We achieve high specificity performance with our condition-limit number MLEM2 with the fewest chemical property features among benchmarked methods. Our method was tested against publicly available prediction servers CAMP AMP prediction [9], iAMP-2 L [50], and a motif-searching algorithm EFC method [51, 52] with and without FCBF. The approach produces physicochemical boundaries that create definitions of similarity among antimicrobial and non-antimicrobial peptides.

Results

The explosion of available antimicrobial peptides brings the new challenge of selecting which antimicrobial peptides to use [38, 56–58]. With the large increase in the number of available peptides, there is an opportunity to classify peptides with respect to their similarity. We define similarity by the physicochemical properties of the peptides, which we show can differentiate between active and inactive peptides. Each rule generated is a category of peptides with boundaries of physicochemical properties chosen so that no rule category is a mixture of active and inactive peptides beyond an allowed limit. We generate rules until all peptides in the training set are covered by at least one category.

Training sets are formatted as data tables; Table 1 is provided as an example to summarize these data sets. The first column is the identity column, which presents the sequences of the peptide. Each row of the data table corresponds to one peptide sequence. The feature columns list the corresponding values for each peptide depending on the amino acid properties and the summarizing function. The final column is the label of antibacterial activity. A condition is a value interval for a feature. The intersection of conditions is a rule, as shown in Fig. 1.

Evaluating the performance of the rules being generated is performed by calculating the *Pr*, the training set accuracy performance of the rule. The *Pr* is the ratio of the size of the sets of peptides described by the intersection of all the conditions in the rule that meet the targeted label to all the peptides described by the intersection of the conditions (Eq. 1). The CLN value is the user-defined condition-limit number, which limits the number of conditions in each of the rules. The value of *Pr* must be at or above α, the user-defined minimum training accuracy a rule must have to be included in the rule set.

Table 1 Schematic Data table representing the training data set before feature correlation analysis. The three sections of the table are the sequences from iAMP-2 L training set [50], the features derived from the 544 amino acid properties in the AAindex1 [63], and the classification label of antibacterial activity from the positive or negative training data set. a_n denotes a sequence, b_n indicates the sum of the sequence for an AAindex1 property, c_n indicates the mean and d_n indicates the maximum sum of three adjacent residues in the sequence

Sequence	Sum of $A_1 \ldots A_{544}$	Mean of $A_1 \ldots A_{544}$	Window of $A_1 \ldots A_{544}$	Antibacterial Activity
a_1	$(b_1)_1 \ldots (b_1)_{544}$	$(c_1)_1 \ldots (c_1)_{544}$	$(d_1)_1 \ldots (d_1)_{544}$	Active
…	…	…	…	Active
a_{1274}	$(b_{1274})_1 \ldots (b_{1274})_{544}$	$(c_{1274})_1 \ldots (c_{1274})_{544}$	$(d_{1274})_1 \ldots (d_{1274})_{544}$	Active
$a_{1,275}$	$(b_{1,275})_1 \ldots (b_{1,275})_{544}$	$(c_{1,275})_1 \ldots (c_{1,275})_{544}$	$(d_{1,275})_1 \ldots (d_{1,275})_{544}$	Inactive
…	…	…	…	Inactive
a_{2714}	$(b_{2714})_1 \ldots (b_{2714})_{544}$	$(c_{2714})_1 \ldots (c_{2714})_{544}$	$(d_{2714})_1 \ldots (d_{2714})_{544}$	Inactive

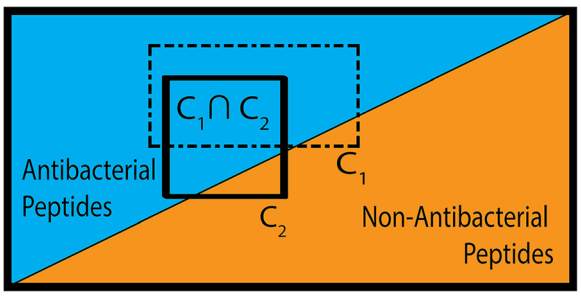

Fig. 1 Rough Set Theory Rule Generation. A) Venn diagram of active and inactive peptides. A rule (R_1) is the intersection of conditions ($C_1$and C_2). Each rule must be selective for either active or inactive peptides. The minimum accuracy allowed for a rule is a user-defined parameter α. B) Venn diagram showing multiple rules as the intersection of conditions in 2-D space. The selection of conditions that lead to rules is a feature selection process that chooses the most relevant conditions to describe the physicochemical boundaries. A rule set is the collection of all rules describing the boundaries for either activity or inactivity

$$Pr = \frac{\left|\cap_1^{CLN} C_i\right|_{targeted\ label}}{\left|\cap_1^{CLN} C_i\right|_{any\ label}} \quad (1)$$

In using the rough set theory approach, we modified existing approaches by combining the features of MLEM2 (modified learning from examples module, Version 2) method [59, 60] with a feature of the module IRIM (Interesting Rule Induction Module) to potentially improve our selectivity and specificity [61]. We modified the MLEM2 method by adding the ability to limit the condition number for each of the rules, a feature of IRIM. Because the IRIM method exhaustively searches all possible rules given the number of conditions, it cannot be used for large numbers of conditions or large numbers of peptides because the runtime grows exponentially with the number of conditions.

Our modified MLEM2 method uses the heuristics of the MLEM2 method to select condition combinations with a run time that grows polynomially in the number of peptides and in the number of conditions. Our modified method includes a defined-condition number (CLN) which combines the polynomially-bound worst-case runtime of MLEM2 with the set number of conditions of IRIM. Because a small number of conditions are selected from the available number of conditions, CLN-MLEM2 is an embedded feature selection method [62]. It attempts to use the most relevant conditions to describe the boundaries. The relevance of a condition is the number of peptides that are described by it in the training set. The CLN-MLEM2 method selects rules based on a user-defined minimum accuracy referred to as α ($0 \le \alpha \le 1$). Using higher values of α generates fewer rules with higher *Pr* values of training accuracy. Using lower values of alpha generates more rules with lower *Pr* values of training accuracy. CLN-MLEM2 generates rules until all peptides in the training set have at least one rule that applies to it. The collection of all rules for either active peptides or inactive peptides is called a rule set.

To begin the defined-condition number MLEM2 (Modified Learning from Experience Module 2) method, we generate multiple summaries of the amino acid sequences of the given active and inactive peptides by selecting non-correlated amino acid properties in the AAindex1 [63] (Amino Acid index 1). Among the 544 properties of the AAindex1, many of the properties are highly correlated. The autocorrelation matrix of the AAindex1 properties was calculated as the pairwise Pearson correlation value of each pair of properties in the index. The heat map of correlation values for the autocorrelation matrix is shown in Fig. 2a. Positive correlation is magenta and negative correlation is teal. Non-correlated amino acid property pairs are white. The autocorrelation matrix shows that most amino acid properties are highly correlated. We studied how many amino acid properties are below a correlation threshold for all other amino acid properties (Fig. 2b). We performed 60 repetitions with random initial properties of eliminating properties more correlated than a threshold. We found a very tight trend of how many uncorrelated properties there are for a given cut-off value. For further study, we selected a correlation cut-off of 0.65, which resulted in 74 properties remaining from the original 544 properties.

We seek to combine overall sequence chemical properties and motif properties to be able to account for how all of the residues may affect the chemical properties while still retaining the ability to separate classifications based on the ordering of the residues. If only chemical properties are evaluated by the sum or mean of the whole sequence, then the rules generated are sequence-order insensitive. By considering sub-sequences of the peptides, then the ordering of the chemical properties within the sequence can be used as a feature. We calculate two types of sequence property summaries from the selected amino acid properties in the AAindex1 (Amino Acid index 1) after removing the correlated amino acid chemical properties. First, we calculate overall property summaries as the mean and average of the properties of the amino acids present in the sequence. Secondly, we calculate motif properties as the maximal subsequence sum of a given length of the amino acid sequence. Our CLN-MLEM2 method can combine overall sequence properties and motif properties within a single rule. Each rule forms a class of either active or inactive peptides.

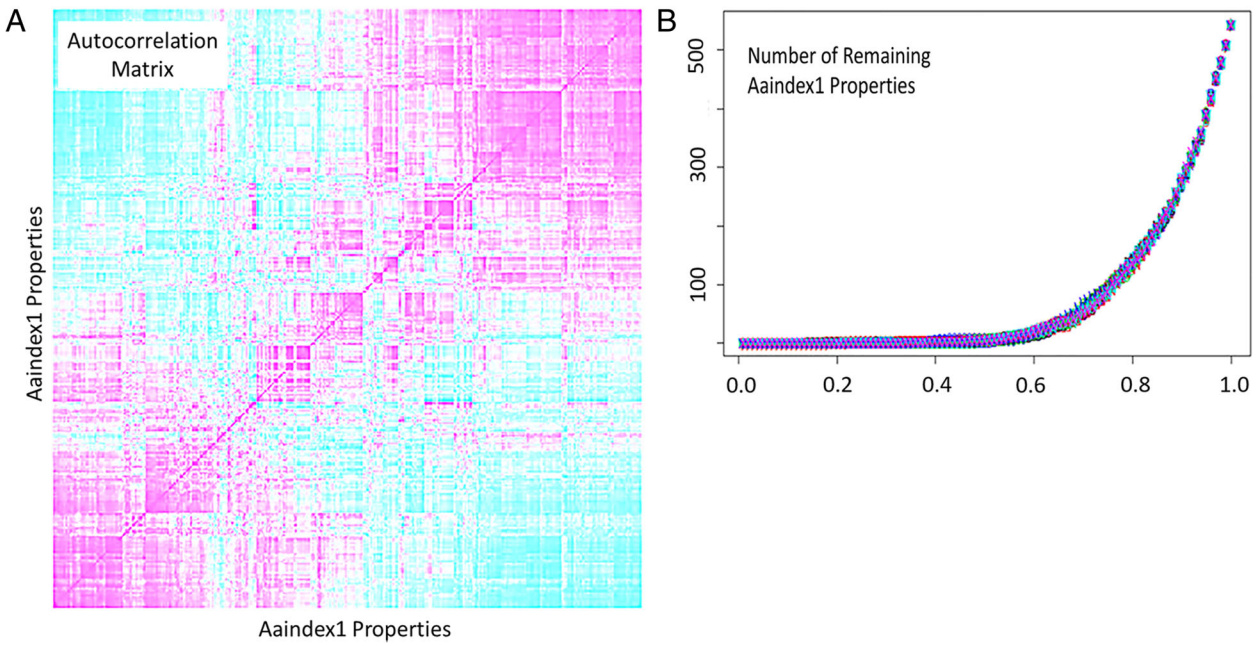

Fig. 2 Auto-Correlation and Selection of AAindex1 Properties. **a** Auto-correlation plot of 544 different AAindex1 properties. Magenta represents positive correlation, cyan represents negative correlation and white represents the lack of correlation between properties. **b** Remaining number of AAindex1 properties following filtering by cut-off value for the absolute value of correlation

We used previously studied, publicly available datasets of antimicrobial peptides [50, 64] to test our method of finding physicochemical boundaries for antibacterial activity. See Table 2 for the inducted rule category with the largest membership of the studied dataset. The rule category is the conjunctive expression of each of the conditions up to the user-defined condition-limit number (CLN) with the rule applying to antimicrobial peptides whose property values are within the range of the values given in Table 2 (Eq. 2). This rule has a high selectivity of 97.8% with a false discovery rate of 2.2%. All sequences that do not match any rule for the applied rule set are classified as non-antibacterial.

$$\bigcap_{1}^{n=CLN}\left(Lower\ Value_{condition} \le Value_{peptide} \le Upper\ Value_{condition}\right) \overset{predicts}{\rightarrow} Antibacterial\ Activity \tag{2}$$

Table 2 Rough set theory rules generated with maximum support from large training dataset. The first rule describes antibacterial sequences. The accuracy of this rule is 97.8% (446/456) for the peptides that met the conditions from either the dataset from Xiao, et al [50] or the dataset from Fernandes, et al. [64]

Calculation	AAindex1 Property	Lower Value	Upper Value
Window 3	NAKH900111	31.21	48.66
Window 3	FINA910104	3.45	5.10
Window 3	KUMS000101	23.6	28.20
Sum	GEIM800102	12.68	39.90
Window 3	VASM830102	1.67	2.12
Window 3	QIAN880139	0.38	0.98
Sum	FAUJ880112	0	3
Sum	CHAM820102	−0.61	19.51

Discussion

Protein and peptide sequence-based classification methods have been extensively developed to improve the understanding of the functionality of polypeptides [65, 66]. By using rough set theory, our method builds rules that distinguish between active antibacterial peptides from inactive antibacterial peptides. The developed method was benchmarked against methods including a recently published method EFC [52], based on motif-recognition, as well as against a larger set of methods from publicly available prediction servers. The first benchmark test is a ten-fold cross validation on a dataset used in previous studies [52, 64] with the positive sequences clustered from the APD2 (Antimicrobial Peptide Database 2) [10] to 115 clusters and the negative sequences from the PDB [67] clustered to 116 clusters. Each cluster is represented by one sequence. The results were compared with EFC-based methods and support vector machines given subsequences of lengths 5 to 8 amino acids. Table 3 demonstrates that

Table 3 Performance of rough set theory rule induction compared to motif-search in 10-fold cross validation

Method	Sensitivity (%)	Specificity (%)	MCC
5-kmer SVM	75.7	75.0	0.54
6-kmer SVM	74.8	74.1	0.46
7-kmer SVM	73.0	72.4	0.40
8-kmer SVM	73.0	72.4	0.36
EFC-FCBF	87.1	87.2	0.76
CLN-MLEM2	86.9	86.3	0.75

our method has high selectivity and accuracy in comparison to the performance of the SVM methods, and comparable selectivity and accuracy in comparison to the EFC method. A trend of decreasing Mathew's Correlation Coefficient (0 for random guessing and 1 for perfect performance) as the length of the subsequence increases is seen in Table 3. Our subsequences in CLN-MLEM2 are 3 amino acids long and may have helped to contribute to our improved performance for using a single length of subsequences instead of combining four different lengths in the EFC method.

We further tested our modified MLEM2 method against a larger variety of classification methods. The second benchmarking test uses the iAMP-2 L dataset [50]. Like the dataset used for the first benchmark, this dataset is derived from the APD2 database. However, instead of choosing a single sequence from each cluster, the sequences were narrowed by removing sequences with greater than 40% similarity as measured by CD-HIT [68] only with cluster of more than 250 sequences. This resulted in a testing positive dataset of 848 unique sequences. The negative sequences were from a UniProt search of cytoplasmic proteins, also with less than 40% similarity. 2405 unique sequences were included in the negative dataset. The positive training data set was the S1 set ("Antibacterial") from iAMP-2 L, which has 1274 unique sequences. The negative training set of data was the non-AMP data set from iAMP-2 L, which has 1440 unique sequences.

While our method has comparable selectivity in classification to current state-of-the-art method, our method is among the best in specificity (Table 4). The combination evolutionary algorithm with chemical properties (EFC + 307-FCBF: EFC combined with FCBF (Fast Correlation Based Features) using 307 features) is the only other state-of-the-art method with specificity that is comparable to ours. We achieve similar specificity using 74 AAindex1 features instead of 307 AAindex1 features. Removing the length-independent representation from the EFC method (EFC-FCBF: EFC without FCBF) results in almost no loss of sensitivity, but a loss of 6% in selectivity. Removing the order-sensitive representation for EFC in Table 2 results in lower sensitivity and selectivity performance (MCC = 0.54). While the datasets are different, between

Table 4 Performance comparison among prediction servers for antimicrobial peptides, a motif-based classification method and rough set theory approach

Method	Sensitivity (%)	Specificity (%)	MCC
CAMP SVM	95.8	39.8	0.43
CAMP RF	97.1	33.5	0.40
CAMP ANN	89.1	70.9	0.61
CAMP DA	94.1	49.5	0.49
iAMP-2 L	97.7	92.0	0.90
EFC-FCBF	92.0	90.0	0.73
EFC + 307-FCBF (307 AAindex1 features)	92.4	96.1	0.86
CLN-MLEM2 (74 AAindex1 features)	88.0	95.4	0.85

Table 3 and Table 4 results, the difference in the individual components of the EFC algorithm compared to the combined algorithm shows a dramatic improvement when integrating order-sensitive and length independent sequence representations. Our CLN-MLEM2 method integrates these two types of representations at its most basic level of output, the rule.

Our method has high specificity and similar accuracy for antibacterial classification as other current methods. When using a classification method for the discovery of antimicrobial peptides, the specificity of the method is more important than its selectivity [69]. Our method prioritizes specificity with low false discovery rate (FDR) by classifying sequences that do not meet any rule in the applied rule set as inactive (Fig. 3). In fact, there is only one method, which provides lower FDR compared to our method, i.e. EFC + 307-FCBF. However, our method results in similar specificity starting with fewer physicochemical properties. The robustness of this method may be potentially improved with ensemble learning and voting scheme approaches. If our method provides unique descriptions of activity, then it will reduce the overall

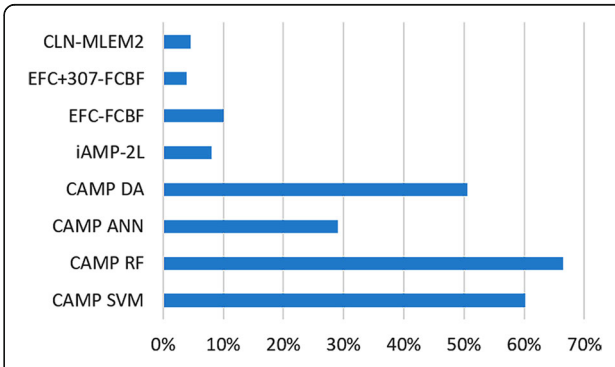

Fig. 3 False discovery rates of comparative antimicrobial peptide classification methods. CLN-MLEM2 achieves a low false discovery rate among currently available antimicrobial peptide classification methods

false discovery rate of the ensemble method and voting scheme approaches.

CLN-MLEM2 has been shown to be useful for the learning task of predicting antibacterial activity from a peptide sequence. This learning task is related to multi-instance learning. A classic literature example of a multi-instance learning problem is in drug activity prediction [70]. Active molecules have at least one conformation that interacts with a drug target, while inactive molecules have none. The challenge is to identify which conformations interact with the drug target. Each drug has one molecular formula, but it can have many conformations. Each peptide also has one sequence but many physicochemical property values. The CLN-MLEM2 method has found the most relevant physicochemical property features that relate to the activity of the peptide sequence. This CLN-MLEM2 method can also be applied to the multi-instance learning case of describing the conformations of peptides are active.

Our method also acts as an embedded feature selection tool by limiting the physicochemical properties in the rules to a user-defined number [62]. This embedded feature selection property may make CLN-MLEM2 useful for feature selection for other methods in the field, with the capability of setting the limit of the number of features to select. Our proposed method, CLN-MLEM2 has a low false discovery rate compared to comparative antimicrobial peptide methods as shown in Fig. 3. EFC method also has a low false discovery rate when including the physicochemical properties, but a doubled false discovery rate when the pattern recognition component is used alone.

A decrease in selectivity of the classification will cause longer computer search times, while a decrease in specificity will increase the number of necessary experimental activity assays. Since the cost of experimentally testing peptides is much greater than the computational time of searching for antimicrobial peptides, methods that have high specificity are preferred. In addition to the high specificity of our method, our method creates categories of antimicrobial peptides. Categorization of peptides aids in the selection and in the design of antimicrobial peptides by providing similarity groupings according to physicochemical property boundaries. Peptides that match multiple active categories can combine more physicochemical property values associated with activity.

Conclusion

The increase in multidrug resistant bacteria usage has prompted an intense search for agents that can be used to treat infectious diseases. There is growing interest in antimicrobial peptides as novel agents to treat infections, and this interest has led to an exponential growth of known antimicrobial peptides. However, peptide selection is becoming another challenge with the drastic increase in the number of these peptides discovered from natural resources, their modified version as well as computational derived ones. We developed a method, CLN-MLEM2, for generating rule sets to describe the similarity among antimicrobial peptides by physicochemical boundaries. Our CLN-MLEM2 method allows the user to limit the number of physicochemical properties used to set the boundaries. Discovering where the boundaries of physicochemical properties are among active peptides generates new categories of antimicrobial peptides.

Our approach simultaneously groups peptides and classifies them. We benchmark our rule set performance to other classification methods. Some available classification methods are either sequence-order insensitive or length-dependent. The rule sets our method generates combine order-sensitive descriptors with length-independent descriptors. We achieve comparable or improved specificity and selectivity to currently available methods with lower false discovery rates. The high specificity of our method aids novel antibacterial peptide discovery because a low false discovery rate reduces the number of bacterial assays.

Methods

In this study we test our rough set theory classification method to differentiate antibacterial peptides from APD2 [10] (Antimicrobial Peptide Database 2) and randomly selected peptides from the UniProt database [71, 72]. These benchmark datasets are available online [50, 64].

Rule induction by the MLEM2 algorithm

The MLEM2 rule induction method [54] is a classification method based on a rough set theory approach that uses local approximations of concepts to generate rules when the available attributes cannot perfectly separate the data. A local approximation is finding collections of conditions that cover a concept with an accuracy requirement parameter α. We use a modified MLEM2 version that combines the polynomial run time growth rate of MLEM2 with the defined-condition number of the IRIM (Interesting Rule Induction Method) to find rules with small numbers of conditions in large datasets with many attributes. IRIM has an exponential run time growth rate with respect to attribute number. We set the maximum number of conditions to be eight (8). Conditions are intervals of feature values. Each peptide sequence has one value for each feature. Rules are conjunctive expressions of conditions.

Figure 4 shows the overall process for building rules. Rules are built from conditions that contain the maximum number of peptide sequence of the desired antibacterial label. Ties are broken by the conditions that have the highest percentage of peptide sequences with

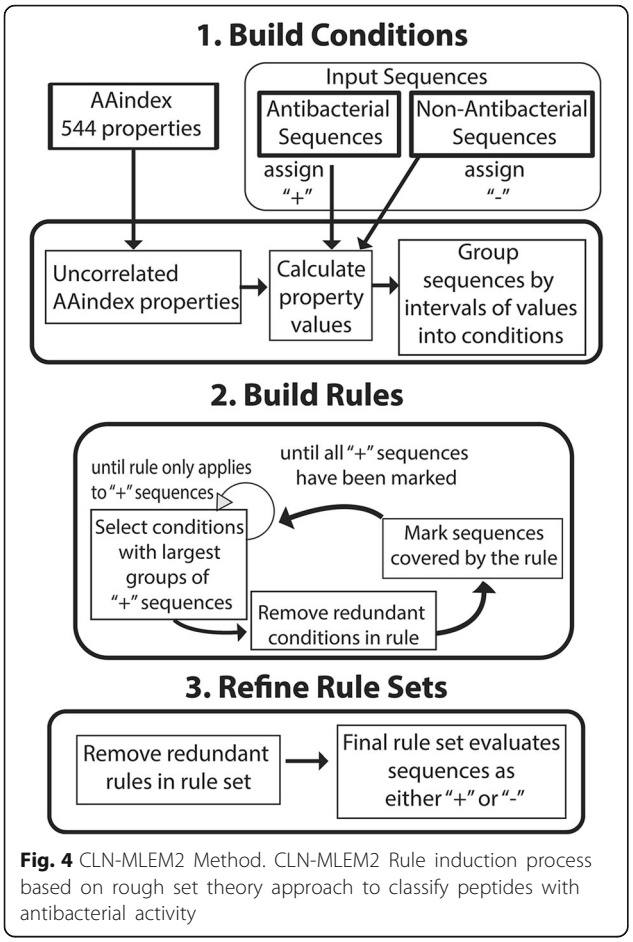

Fig. 4 CLN-MLEM2 Method. CLN-MLEM2 Rule induction process based on rough set theory approach to classify peptides with antibacterial activity

the desired antibacterial label. Rules are refined by narrowing the interval of an included condition or by adding a new condition to the conjunctive expression. Rules are simplified by omitting redundant conditions whose loss still results in a rule with no loss of accuracy. The minimum accuracy that a valid rule must have is a user-defined value, α. In this study, α is set to the accuracy of the majority class rule, which is to label all peptides with the non-antibacterial class.

Table 5 shows a compact data table that is consisted of six sequences with two features to illustrate methodology. The most relevant condition among the two features for active antibacterial activity is the sum of the positive charge from 1 to 3, relating to all three active peptides. This condition does not form a rule, however there is an inactive sequence with a sum of positive charge of 1. To distinguish between inactive and active between these two sequences, the second feature of the sum of negative charges is considered. The intersection of the conditions of the sum of positive charge from 1 to 3 and the sum of negative charge from 0 to 1 is a valid rule for labeling active peptides for this data table. This rule forms a boundary between active and inactive peptides for this data table. In larger data tables, rules may also form boundaries between active peptides or between inactive peptides because different features may be relevant for the activity for different sets of peptides.

Table 5 Data table consists of six selected sequences with two features

Sequence	Sum of Sum of FAUJ880111	Sum of Sum of FAUJ880112	Antibacterial Activity
FFPVIGRILNGIL	1	0	Active
KFHEKHHSHRGY	3	1	Active
GNNRPVYIPQPRPPHPRL	3	0	Active
QDVDHVFLRF	1	2	Inactive
QQDYTGWMDF	0	1	Inactive
QLTFTSSWG	0	0	Inactive

Correlated AAindex1 property removal

The AAindex1 has 544 properties with one value for each of the twenty naturally occurring amino acids [63]. A database of all properties is available in the R package 'seqinr' [73]. We constructed an autocorrelation matrix of these properties to provide pairwise correlation comparisons for all 544 properties. We filtered properties using an absolute correlation value cutoff. We randomized which property to keep by randomizing the order in which the properties were compared.

Performance descriptions

In binary classification there are two different descriptions of performance based on the two possible error types, false positives and false negatives. Sensitivity refers to the likelihood of correctly predicting a positive result, while specificity refers to the likelihood of correctly predicting a negative result. Sensitivity deals with avoiding false positives, while specificity deals with avoiding false negatives. Selectivity, which can be directly derived from specificity, is the likelihood of incorrectly predicting a negative result, a false negative. Further details about performance measures are included in Additional file 1.

Additional file

Additional file 1: Feature Generation and Performance Measure Methods (DOCX 30 kb)

Abbreviations

AAindex1: Amino acid index 1; AMP: Antimicrobial peptide; ANN: Artificial neural network; APD2: Antimicrobial peptide database 2; CAMP: Collections of antimicrobial peptides; CLN: Condition limit number; DA: Discriminant analysis; EFC + 307-FCBF: Evolutionary feature construction and fast correlation-based filter selection with 307 features; EFC: Evolutionary feature construction; EFC-FBCF: Evolutionary feature construction without fast correlation-based filter selection; FBCF: Fast correlation-based filter selection; FDR: False discovery rate; FN: False negative; FP: False positive; HMM: Hidden Markov model; iAMP-2 L: Antimicrobial peptide prediction two-level;

IRIM: Interesting rule induction method; LR: Logistic regression; MCC: Matthew's correlation coefficient; MLEM2: Modified Learning from Experience Module 2; QM: Quantitative matrix; SVM: Support vector machine; TN: True negative; TP: True positive

Acknowledgements
We acknowledge the valuable scientific discussions with Professor Malcolm L. Snead (University of Southern California) to address the challenges and the opportunities on antimicrobial peptide design. We are also thankful to Cate E. Wisdom for her ongoing support on antimicrobial peptide studies to test and characterize the functions of the peptides.

Funding
This investigation was supported by research grants R01DE022054, 3R01DE022054-04S1 and R01DE025476 from the National Institute of Dental and Craniofacial Research, and from National Institute of Arthritis and Musculoskeletal and Skin Diseases R21AR062249, National Institutes of Health, Bethesda, Maryland. The funding sources had no role in any of the following: the design of the study, the collection of data, the analysis of data, or the interpretation of data.

Availability of data and materials
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Authors' contributions
KB developed the theory, performed the computations and wrote the initial manuscript. KC contributed the design, analysis and verification of data. PS contributed to analyses of the data and the scientific content. CT initiated the topic of antimicrobial peptide study, conceived and supervised the work. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Bioengineering Program, Institute of Bioengineering Research, University of Kansas, Learned Hall, Room 5109, 1530 W 15th Street, Lawrence, KS 66045, USA. [2]Chemical and Petroleum Engineering Department, University of Kansas, Learned Hall, Room 4154, 1530 West 15th Street, Lawrence, KS 66045, USA. [3]Mechanical Engineering Department, Bioengineering Program, Institute of Bioengineering Research, University of Kansas, Learned Hall, Room 3111, 1530 West 15th Street, Lawrence, KS 66045, USA. [4]Mechanical Engineering Department, Bioengineering Program, Institute of Bioengineering Research, University of Kansas, Learned Hall, Room 3135A, 1530 W 15th St, Lawrence, KS 66045, USA.

Received: 28 June 2018 Accepted: 20 November 2018
Published online: 06 December 2018

References

1. Ventola CL. The antibiotic resistance crisis: part 1: causes and threats. Pharmacy and Therapeutics. 2015;40(4):277–83.
2. Mishra B, Reiling S, Zarena D, Wang G. Host defense antimicrobial peptides as antibiotics: design and application strategies. Curr Opin Chem Biol. 2017;38:87–96.
3. Piddock L. Reflecting on the final report of the O'Neill Review on Antimicrobial Resistance. Lancet Infect Dis. 2016;767–68. https://doi.org/10.1016/S1473-3099(16)30127-X
4. Al-Tawfiq JA, Laxminarayan R, Mendelson M. How should we respond to the emergence of plasmid-mediated colistin resistance in humans and animals? Int J Infect Dis. 2017;54:77–84.
5. Fan L, Sun J, Zhou M, Zhou J, Lao X, Zheng H, Xu H. DRAMP: a comprehensive data repository of antimicrobial peptides. Sci Rep. 2016;6:24482.
6. Di Luca M, Maccari G, Maisetta G, Batoni G. BaAMPs: the database of biofilm-active antimicrobial peptides. Biofouling. 2015;31(2):193–9.
7. Wang G, Li X, Wang Z. APD3: the antimicrobial peptide database as a tool for research and education. Nucleic Acids Res. 2016;44(D1):D1087–93.
8. Zhao X, Wu H, Lu H, Li G, Huang Q. LAMP: a database linking antimicrobial peptides. PLoS One. 2013;8(6):e66557.
9. Thomas S, Karnik S, Barai RS, Jayaraman VK, Idicula-Thomas S. CAMP: a useful resource for research on antimicrobial peptides. Nucleic Acids Res. 2010;38(Database issue):D774–80.
10. Wang G, Li X, Wang Z. APD2: the updated antimicrobial peptide database and its application in peptide design. Nucleic Acids Res. 2009;37(Database issue):D933–7.
11. Wang J, Dong XQ, Yu QS, Balzer SN, Li H, Larm NE, Balzer GA, Chen L, Tan JW, Chen M. Incorporation of antibacterial agent derived deep eutectic solvent into an active dental composite. Dent Mater. 2017;33(12):1445–55.
12. Chen YX, Mant CT, Farmer SW, Hancock REW, Vasil ML, Hodges RS. Rational design of alpha-helical antimicrobial peptides with enhanced activities and specificity/therapeutic index. J Biol Chem. 2005;280(13):12316–29.
13. Wisdom C, VanOosten SK, Boone KW, Khvostenko D, Arnold PM, Snead ML, Tamerler C. Controlling the biomimetic implant interface: modulating antimicrobial activity by spacer design. J Mol Eng Mater. 2016;4(1):1640005.
14. Yazici H, ONeill MB, Kacar T, Wilson BR, Oren EE, Sarikaya M, Tamerler C. Engineered chimeric peptides as antimicrobial surface coating agents towards infection-free implants. ACS Appl Mater Interfaces. 2016;8(8):5070–81.
15. Yucesoy DT, Hnilova M, Boone K, Arnold PM, Snead ML, Tamerler C. Chimeric peptides as implant functionalization agents for titanium alloy implants with antimicrobial properties. JOM. 2015;67(4):754–66.
16. Tajbakhsh M, Karimi A, Tohidpour A, Abbasi N, Fallah F, Akhavan MM. The antimicrobial potential of a new derivative of cathelicidin from Bungarus fasciatus against methicillin-resistant Staphylococcus aureus. J Microbiol. 2018;56(2):128–37.
17. Vasudev PG, Chatterjee S, Shamala N, Balaram P. Structural chemistry of peptides containing backbone expanded amino acid residues: conformational features of beta, gamma. and Hybrid Peptides, Chemical reviews. 2011;111(2):657–87.
18. Sang P, Shi Y, Teng P, Cao AN, Xu H, Li Q, Cai JF. Antimicrobial AApeptides. Curr Top Med Chem. 2017;17(11):1266–79.
19. Seebach D, Beck AK, Bierbaum DJ. The world of beta- and gamma-peptides comprised of homologated proteinogenic amino acids and other components. Chem Biodivers. 2004;1(8):1111–239.
20. Porter EA, Weisblum B, Gellman SH. Mimicry of host-defense peptides by unnatural oligomers: antimicrobial beta-peptides. J Am Chem Soc. 2002; 124(25):7324–30.
21. Knerr PJ, van der Donk WA. Discovery, Biosynthesis, and Engineering of Lantipeptides. In: Kornberg RD, editor. Annual Review of Biochemistry, vol. 812012. p. 479–505.
22. Brogden NK, Brogden KA. Will new generations of modified antimicrobial peptides improve their potential as pharmaceuticals? Int J Antimicrob Agents. 2011;38(3):217–25.
23. Candido-Ferreira IL, Kronenberger T, Sayegh RSR, Batista IDC, da Silva PI. Evidence of an antimicrobial peptide signature encrypted in HECT E3 ubiquitin ligases. Front Immunol. 2017;7.
24. Loose C, Jensen K, Rigoutsos I, Stephanopoulos G. A linguistic model for the rational design of antimicrobial peptides. Nature. 2006;443(7113):867–9.
25. Porto WF, Pires AS, Franco OL. Computational tools for exploring sequence databases as a resource for antimicrobial peptides. Biotechnol Adv. 2017; 35(3):337–49.
26. Boone K, Abedin F, Anwar MR, Camarda KV. Molecular Design in the Pharmaceutical Industries. Computer Aided Chemical Engineering. 2017; 39:221–38.
27. Ng LY, Chong FK, Chemmangattuvalappil NG. Challenges and opportunities in computer-aided molecular design. Comput Chem Eng. 2015;81:115–29.
28. Roughton BC, Christian B, White J, Camarda KV, Gani R. Simultaneous design of ionic liquid entrainers and energy efficient azeotropic separation processes. Comput Chem Eng. 2012;42:248–62.

29. Lin B, Chavali S, Camarda K, Miller DC. Computer-aided molecular design using Tabu search. Comput Chem Eng. 2005;29(2):337–47.
30. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, Dearden J, Gramatica P, Martin YC, Todeschini R, Consonni V, Kuz'min VE, Cramer R, Benigni R, Yang C, Rathman J, Terfloth L, Gasteiger J, Richard A, Tropsha A. QSAR modeling: where have you been? Where are you going to? J Med Chem. 2014;57(12):4977–5010.
31. Riera-Fernandez P, Martin-Romalde R, J. Prado-Prado F, Escobar M, R. Munteanu C, Concu R, Duardo-Sanchez A, Gonzalez-Diaz H. From QSAR models of drugs to complex networks: state-of-art review and introduction of new Markov-spectral moments indices. Curr Top Med Chem. 2012;12(8):927–60.
32. Prado-Prado FJ, Uriarte E, Borges F, Gonzalez-Diaz H. Multi- target spectral moments for QSAR and complex networks study of antibacterial drugs. Eur J Med Chem. 2009;44(11):4516–21.
33. Du Q-S, Huang R-B, Chou K-C. Recent advances in QSAR and their applications in predicting the activities of chemical molecules, peptides and proteins for drug design. Current protein and peptide science. 2008;9(3):248–59.
34. Fjell CD, Hiss JA, Hancock RE, Schneider G. Designing antimicrobial peptides: form follows function. Nat Rev Drug Discov. 2011;11(1):37–51.
35. Fjell CD, Jenssen H, Cheung WA, Hancock RE, Cherkasov A. Optimization of antibacterial peptides by genetic algorithms and cheminformatics. Chem Biol Drug Des. 2011;77(1):48–56.
36. Cherkasov A, Hilpert K, Jenssen H, Fjell CD, Waldbrook M, Mullaly SC, Volkmer R, Hancock REW. Use of artificial intelligence in the Design of Small Peptide Antibiotics Effective against a broad Spectrum of highly antibiotic-resistant superbugs. ACS Chem Biol. 2009;4(1):65–74.
37. Claro B, Bastos M, Garcia-Fandino R. Design and applications of cyclic peptides. In: Peptide Applications in Biomedicine, Biotechnology and Bioengineering. Cambridge: Woodhead Publishing, Elsevier; 2018. pp. 87–129. https://doi.org/10.1016/B978-0-08-100736-5.00004-1.
38. Muller AT, Kaymaz AC, Gabernet G, Posselt G, Wessler S, Hiss JA, Schneider G. Sparse neural network models of antimicrobial peptide-activity relationships. Molecular Informatics. 2016;35(11–12):606–14.
39. Gabere MN, Noble WS. Empirical comparison of web-based antimicrobial peptide prediction tools. Bioinformatics. 2017;33(13):1921–9.
40. Lata S, Sharma BK, Raghava GPS. Analysis and prediction of antibacterial peptides. Bmc Bioinformatics. 2007;8.
41. Bhasin M, Raghava GP. SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence. Bioinformatics. 2004;20(3):421–3.
42. Bhasin M, Raghava GPS. Prediction of CTL epitopes using QM. SVM and ANN techniques, Vaccine. 2004;22(23):3195–204.
43. Saha S, Raghava G. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. Proteins: Structure, Function, and Bioinformatics. 2006;65(1):40–8.
44. Waghu FH, Gopi L, Barai RS, Ramteke P, Nizami B, Idicula-Thomas S. CAMP: Collection of sequences and structures of antimicrobial peptides. Nucleic Acids Res. 2014;42(1):D1154–D1158.
45. Karatzoglou A, Smola A, Hornik K, Zeileis A. kernlab-an S4 package for kernel methods in R. J Stat Softw. 2004;11(9):1–20.
46. Venerables W, Ripley B. Modern applied statistics with S. New York: Springer; 2002.
47. Bhadra P, Yan J, Li J, Fong S, Siu SWI. AmPEP: sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest. Sci Rep. 2018;8:1697.
48. Norušis MJ. SPSS/PC+ advanced statistics V2. 0: for the IBM PC/XT/AT and PS/2, SPSS Incorporated; 1988.
49. Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002;2(3):18–22.
50. Xiao X, Wang P, Lin W-Z, Jia J-H, Chou K-C. iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. Anal Biochem. 2013;436(2):168–77.
51. Meher PK, Sahu TK, Saini V, Rao AR. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. Sci Rep. 2017;7:42362.
52. Veltri D, Kamath U, Shehu A. A novel method to improve recognition of antimicrobial peptides through distal sequence-based features. IEEE International Conference on Bioinformatics and Biomedicine. 2014:371–8.
53. Veltri D, Kamath U, Shehu A. Improving recognition of antimicrobial peptides and target selectivity through machine learning and genetic programming. Ieee-Acm Transactions on Computational Biology and Bioinformatics. 2017;14(2):300–13.
54. Hiroshi S, Wu M, Nakata M. Apriori-based rule generation in incomplete information databases and non-deterministic information systems. Fundamenta Informaticae. 2014;130(3):343–76.
55. Grzymala-Busse JW, Hamilton J, Hippe ZS. Diagnosis of melanoma using IRIM, a data mining system. In: Artificial Intelligence and Soft Computing. Berlin: Springer; 2004;3070:996–1001. https://doi.org/10.1007/978-3-540-24844-6_155.
56. Yu D, Sheng ZG, Xu XQ, Li JX, Yang HL, Liu ZG, Rees HH, Lai R. A novel antimicrobial peptide from salivary glands of the hard tick. Ixodes sinensis, Peptides. 2006;27(1):31–5.
57. Wang GS, Watson KM, Peterkofsky A, Buckheit RW. Identification of novel human immunodeficiency virus type 1-inhibitory peptides based on the antimicrobial peptide database. Antimicrob Agents Chemother. 2010;54(3):1343–6.
58. Menousek J, Mishra B, Hanke ML, Heim CE, Kielian T, Wang GS. Database screening and in vivo efficacy of antimicrobial peptides against methicillin-resistant Staphylococcus aureus USA300. Int J Antimicrob Agents. 2012;39(5):402–6.
59. Grzymala-Busse JW, Rzasa W. A local version of the MLEM2 algorithm for rule induction. Fundamenta Informaticae. 2010;100(1):99–116.
60. Clark PG, Gao C, Grzymala-Busse JW. Complexity of rule sets induced by two versions of the MLEM2 rule induction algorithm. In: Artificial Intelligence and Soft Computing. Cham: Springer; 2017. pp. 21–30. https://doi.org/10.1007/978-3-319-59060-8_3
61. Grzymala-Busse JW, Hamilton J, Hippe ZS. Diagnosis of melanoma using IRIM, a data mining system. In: Rutkowski L, Siekmann JH, Tadeusiewicz R, Zadeh LA, editors. Artificial intelligence and soft computing - ICAISC 2004: 7th international conference, Zakopane, Poland, June 7–11, vol. 2004. Berlin, Heidelberg: Proceedings, Springer Berlin Heidelberg; 2004. p. 996–1001.
62. Austin ND, Sahinidis NV, Trahan DW. Computer-aided molecular design: an introduction and review of tools, applications, and solution techniques. Chem Eng Res Des. 2016;116:2–26.
63. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res. 2008;36(Database issue):D202–5.
64. Fernandes FC, Rigden DJ, Franco OL. Prediction of antimicrobial peptides based on the adaptive neuro-fuzzy inference system application. Biopolymers. 2012;98(4):280–7.
65. Meher P, Sahu T, Gahoi S, Rao A. ir-HSP: Improved recognition of heat shock proteins, their families and sub-types based on g-spaced di-peptide features and support vector machine. Front Genet. 2018;8:235. https://doi.org/10.3389/fgene.
66. Sharma R, Bayarjargal M, Tsunoda T, Patil A, Sharma A. MoRFPred-plus: computational identification of MoRFs in protein sequences using physicochemical properties and HMM profiles. J Theor Biol. 2018;437:9–16.
67. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res. 2000;28(1):235–42.
68. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658–9.
69. Porto WF, Pires AS, Franco OL. Antimicrobial activity predictors benchmarking analysis using shuffled and designed synthetic peptides. J Theor Biol. 2017;426:96–103.
70. Dietterich TG, Lathrop RH, Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles. Artif Intell. 1997;89(1–2):31–71.
71. Magrane M, UniProt C. UniProt knowledgebase: a hub of integrated protein data. Database (Oxford). 2011;2011:bar009.
72. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. The universal protein resource (UniProt). Nucleic Acids Res. 2005;33(Database issue):D154–9.
73. Charif D, Lobry JR. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Structural approaches to sequence evolution. Berlin: Springer; 2007. pp. 207–32. https://doi.org/10.1007/978-3-540-35306-5_10