BioData Mining

METHODOLOGY Open Access

Application of an interpretable classification model on Early Folding Residues during protein folding

Sebastian Bittrich[1,2*†], Marika Kaden[1†], Christoph Leberecht[1,2], Florian Kaiser[1,2], Thomas Villmann[1] and Dirk Labudde[1]

*Correspondence:
sebastian.bittrich@hs-mittweida.de
†Sebastian Bittrich and Marika Kaden contributed equally to this work.
[1]University of Applied Sciences Mittweida, Technikumplatz 17, 09648 Mittweida, Germany
[2]Biotechnology Center (BIOTEC) TU Dresden, Tatzberg 47/49, 01307 Dresden, Germany

Abstract

Background: Machine learning strategies are prominent tools for data analysis. Especially in life sciences, they have become increasingly important to handle the growing datasets collected by the scientific community. Meanwhile, algorithms improve in performance, but also gain complexity, and tend to neglect interpretability and comprehensiveness of the resulting models.

Results: Generalized Matrix Learning Vector Quantization (GMLVQ) is a supervised, prototype-based machine learning method and provides comprehensive visualization capabilities not present in other classifiers which allow for a fine-grained interpretation of the data. In contrast to commonly used machine learning strategies, GMLVQ is well-suited for imbalanced classification problems which are frequent in life sciences. We present a Weka plug-in implementing GMLVQ. The feasibility of GMLVQ is demonstrated on a dataset of Early Folding Residues (EFR) that have been shown to initiate and guide the protein folding process. Using 27 features, an area under the receiver operating characteristic of 76.6% was achieved which is comparable to other state-of-the-art classifiers. The obtained model is accessible at https://biosciences.hs-mittweida.de/efpred/.

Conclusions: The application on EFR prediction demonstrates how an easy interpretation of classification models can promote the comprehension of biological mechanisms. The results shed light on the special features of EFR which were reported as most influential for the classification: EFR are embedded in ordered secondary structure elements and they participate in networks of hydrophobic residues. Visualization capabilities of GMLVQ are presented as we demonstrate how to interpret the results.

Keywords: Machine learning, Visualization, Protein folding, Early folding residues, Residue graphs, Learning vector quantization, Interpretable models

Introduction

The analysis of data collected during biological experiments poses a challenge for modern bioinformatics. Usually this data is feature rich, yet hard to interpret, such as it is the case for single-cell gene expression data obtained by high-throughput experiments [1]. Despite sophisticated pre-processing and the application of machine learning models, analysis – and most importantly interpretation – of such data is still hard to accomplish. Nevertheless, machine learning is the basis for sophisticated predictions and allows new insights into open questions. In this paper, we examine the problem of protein folding, by means of Early Folding Residues (EFR). Further, we apply an interpretable classifier on this problem to deepen the understanding of EFR based on the trained model. We demonstrate how this sparse classification model can be readily discussed and want to sensitize users that this degree of interpretability – though valuable to gain biological insights – is not achievable by most state-of-the-art machine learning techniques.

Grasping the protein folding problem through Early Folding Residues

Proteins are chains of amino acids which are connected by covalent bonds and, for the most part, autonomously fold into a defined structure (Fig. 1) [2, 3]. This stable, three-dimensional structure allows proteins to be functional and catalyze particular chemical reactions, transport molecules, or transduce signals in cells. The fundamentals of the so-called protein folding process are still unclear.

Folding intermediates are highly unstable and thus protein folding was difficult to investigate experimentally for a long time [4, 5]. Nowadays, pulse labeling hydrogen-deuterium exchange is a prominent tool to investigate the folding process with spatial and temporal resolution [6–13]. EFR were identified as key residues for the folding process as they participate in the earliest folding events. By forming long-range tertiary contacts, EFR are

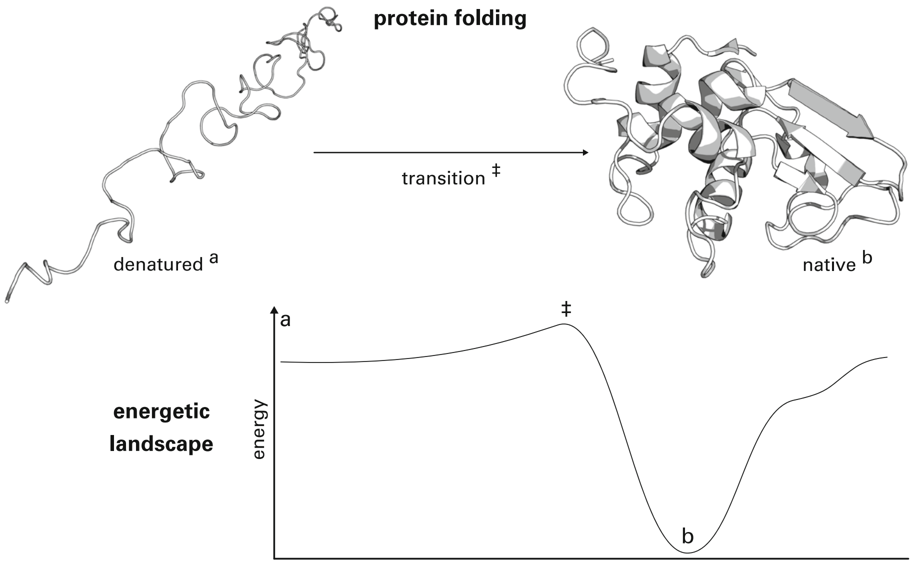


Fig. 1 Illustration of the protein folding process. The denatured protein (**a**) has to pass an energetic barrier (‡), the so-called transition state, to reach its native three-dimensional structure (**b**). Usually, the native structure represents the global energetic optimum of the protein. EFR are residues which initiate and guide the folding process [13, 19]

also assumed to guide the assembly of different protein regions which stabilize the native, folded protein structure [10–13]. EFR were shown to be the initiators of the folding process and, thus, may be used to advance the understanding of the protein folding process [14–17]. The process remains elusive, but understanding the peculiarities of the folding nucleus [12, 16–19] (as indicated by EFR) may aid unraveling it by providing information on intermediate states of the process. The currently identified set of EFR is deposited in the Start2Fold database [13] which provides a robust dataset for the characterization of EFR and the design of classifiers for their prediction. Raimondi et al. developed the predictor *EFoldMine* that discriminates EFR from other residues, termed Late Folding Residues (LFR), using features derived from the protein sequence [16]. The aim of their paper was to distinguish between these two classes using secondary structure propensities and backbone rigidity values of surrounding sequence fragments. It is crucial to understand, which features cause a small number of residues to become EFR while the majority of them are LFR. Unfortunately, the classifiers applied by Raimondi et al. [16] cannot provide detailed insights and the published model is not discussed under this focus. This is mainly the consequence of the chosen features and the employed standard Support Vector Machine (SVM) with the Radial Basis Function (RBF)-kernel; this results in a model which is difficult to interpret and does not state the features relevant to distinguish the classes.

We created a dataset using the same data basis but utilize a more diverse set of features. This set includes information derived from the protein structure as a corresponding structure is deposited in the Protein Data Bank (PDB) for each protein of the dataset. This allows for a better interpretability and discussion of the resulting model and, thus, emphasizes unique aspects of the Generalized Matrix Learning Vector Quantization (GMLVQ) classifier. Our study demonstrates how an adaptation of an established machine learning strategy allows pinpointing the most influential features for classification. Therefore, we present a novel implementation of the GMLVQ algorithm [20, 21] as plug-in for the popular Waikato Environment For Knowledge Analysis (Weka) framework [22–24]. This plug-in features diverse visualization tools which encourage the user to interpret the resulting model and render GMLVQ a comprehensible *white box* classifier. Furthermore, Weka allows to readily access the trained model by the provided application programming interface. Since user-friendly and publicly accessible web applications represent the future direction of the development of machine learning models [25–28], we deployed our model as web server accessible at https://biosciences.hs-mittweida.de/efpred/. The web server displays the predicted positions of EFR in the structure using NGL [29, 30].

Detailed description of the dataset of Early Folding Residues

The Start2Fold database [13, 19] contains the results of pulse labeling hydrogen-deuterium exchange experiments for 30 proteins. Due to the nature of the experimental data, no information can be obtained for the amino acid proline because its amide group is not susceptible to an exchange of hydrogen to deuterium. We extracted annotations of EFR from this database as described by [19] to compose our dataset. All 111 proline instances were dropped from the initial dataset which resulted in 3266 residues of which 482 (14.8%) are EFR. The experimental annotation of a residue to be either EFR or LFR was assigned as class label.

Feature annotation

Every amino acid in the dataset was represented by a number of features capturing different aspects of their molecular surroundings and physicochemical properties. Amino acids have sequential and spatial neighbors and both levels of organization are strongly intertwined by the process of protein folding [31]. All considered features describe a particularized aspect of this connection and are summarized in Table 1. Features of each residue were averaged with respect to four adjacent positions at the sequence level in N- as well as C-terminal region. The dataset is provided in Additional files 1 and 2. Additional file 3 captures correlations between features.

Energy profiling Energy Profiles [32, 33] transform the three-dimensional arrangement of atoms in a protein into a vector of energy values describing each amino acid. The computed energy (*e*) of a residue describes its interactions with its surroundings. Energy Profiles can also be predicted using only sequence information [32] (*ePred*) which represents the sequence composition. Computed as well as predicted energy values have

Table 1 Denomination and short description of the 27 features of the dataset for individual residues classification

Feature	Description
e	Computed energy values
ePred	Predicted energy values
SecSize	Size of the surrounding secondary structure elements
LF	Fraction of surrounding unordered secondary structure elements
Rasa	Relative accessible surface area
PlipLC	Absolute count of local PLIP contacts
PlipHbLC	Absolute count of local PLIP hydrogen bonds
PlipHpLC	Absolute count of local PLIP hydrophobic interactions
PlipBbLC	Absolute count of local PLIP backbone contacts
PlipLR	Absolute count of long-range PLIP contacts
PlipHbLR	Absolute count of long-range PLIP hydrogen bonds
PlipHpLR	Absolute count of long-range PLIP hydrophobic interactions
PlipBbLR	Absolute count of long-range PLIP backbone contacts
PlipBN	Betweenness using all PLIP contacts
PlipCL	Closeness using all PLIP contacts
PlipCC	Clustering coefficient using all PLIP contacts
PlipHbBN	Betweenness using PLIP hydrogen bonds
PlipHbCL	Closeness using PLIP hydrogen bonds
PlipHbCC	Clustering coefficient using PLIP hydrogen bonds
PlipHpBN	Betweenness using PLIP hydrophobic interactions
PlipHpCL	Closeness using PLIP hydrophobic interactions
PlipHpCC	Clustering coefficient using PLIP hydrophobic interactions
ConvBN	Betweenness using the distance-based contact definition
ConvCL	Closeness using the distance-based contact definition
ConvCC	Clustering coefficient using the distance-based contact definition
PlipNC	Distinct neighborhood count using all PLIP contacts
ConvNC	Distinct neighborhood count using the distance-based contact definition

References to these features are given in *italic* font been used before for the description of the folding process [32] as well as protein structure quality assessment [33].

Secondary structure elements Secondary structure elements were annotated using DSSP [34] in its BioJava [35, 36] implementation. The secondary structure element size of a residue (*SecSize*) refers to the number of sequence neighbors sharing the same secondary structure (i.e. α-helix, β-strand, and coil). For sequence windows of nine residues the number of unordered secondary structure elements was counted and normalized by the window size [37]. This yields a fraction (*LF*), where high values are tied to regions of high disorder, whereas amino acids embedded in α-helices or β-sheets result in scores close to 0.

Relative accessible surface area The Relative Accessible Surface Area (RASA) of a residue describes how exposed it is towards to solvent. Residues in the hydrophobic core tend to be buried and exhibit no accessible surface area. RASA values (*Rasa*) were computed with the BioJava [35, 36] implementation of the algorithm by Shrake and Rupley [38].

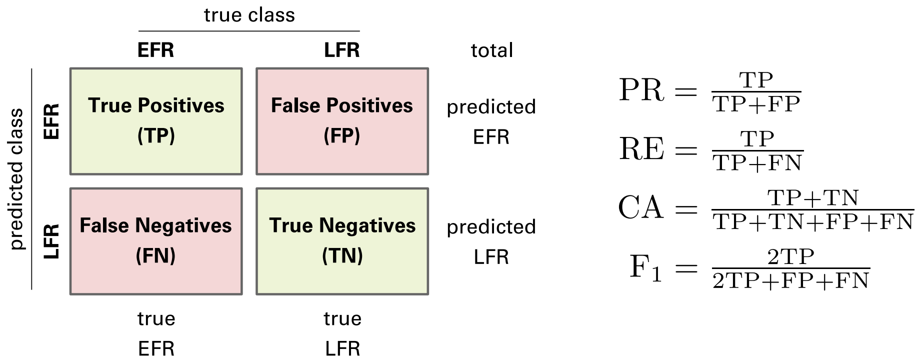
Non-covalent contacts Non-covalent contacts stabilize protein structures and are the driving force behind protein folding [31]. The Protein-Ligand Interaction Profiler (PLIP) [39] was used for the annotation of non-covalent contacts between residues in protein structures. PLIP supports different contact types such as salt bridges, π-stacking interactions, or π-cation interactions. For this study, only hydrogen bonds (*Hb*) and hydrophobic interactions (*Hp*) were considered explicitly. Other contact types were not observed for most of the rather small proteins in the dataset. Furthermore, local and long-range contacts [40] were distinguished. Local contacts (suffix *LC*) are defined as contacts between residues less than six sequence positions apart – their main contribution is stabilizing secondary structure elements. In contrast, long-range contacts (suffix *LR*) occur between residues more than five sequence positions apart and constitute stabilizing contacts between secondary structure elements which primarily manifest the three-dimensional arrangement of a protein. Backbone contacts (*Bb*) occur only between backbone atoms of the respective residues.

Residue graph representation of proteins Proteins in the dataset were represented as residue graphs. Amino acids always constituted the nodes and contacts between residues were represented by edges. Covalently bound residues were considered to be in contact. All contacts annotated by PLIP were used to create the first graph representation (using the prefix *Plip*). Reduced representations were created by only considering hydrogen bonds (prefix *PlipHb*) respectively hydrophobic interactions (prefix *PlipHp*). The contacts detected by PLIP may ignore spatially close residues when they do not form any contacts according the underlying rule set. Therefore, an additional contact definition was employed (prefix *Conv*): two residues were considered to be in contact, if their C_α atoms were at most 8 Å apart.

Topological descriptors Based on the four graph representations (*Plip*, *PlipHb*, *PlipHp*, and *Conv*), topological descriptors of individual residues were computed. This allows to describe how residues are connected to other residues. Most of these properties are based on shortest paths observable in the graph. The betweenness centrality (BN) of a node is defined as the number of shortest paths passing through that particular node. The term is normalized by the number of node pairs $0.5 \cdot n \cdot (n - 1)$ in the residue graph with n nodes [41, 42]. The closeness centrality (CL) of a node is defined the inverse of the average path length to any other node. The clustering coefficient describes the surroundings of individual nodes. All adjacent nodes are collected and the number of edges between these n_k nodes is determined. The clustering coefficient (CC) of a node is defined as number of edges between its adjacent nodes, divided by the maximum number of edges which can theoretically connect these nodes which is $0.5 \cdot n_k \cdot (n_k - 1)$. The distinct neighborhood count (NC) captures how many sequentially distant (long-range) protein regions are connected by a residue [17].

Description of the Generalized Matrix Learning Vector Quantization classifier

The Generalized Learning Vector Quantization (GLVQ) is a powerful distance- and prototype-based classification method [20]. The idea is adapted from the unsupervised vector quantization methods such as k-Means or the Self-Organizing Map (SOM) and an extension of the heuristic Learning Vector Quantization (LVQ) [43]. For each class at least one prototype is initialized and a function, which approximates the classification accuracy (Fig. 2), is maximized during learning. The optimization is commonly done by Stochastic Gradient Ascent (SGA) and ends up in an intuitive adaption of the prototypes. Thereby, in each iteration, for one training data point $\mathbf{v}$ two prototypes are taken into account: the nearest prototype with the same label as the data point and the nearest prototype with a different label, noted as $\mathbf{w}^+(\mathbf{v})$ and $\mathbf{w}^-(\mathbf{v})$. The prototype $\mathbf{w}^+(\mathbf{v})$ is attracted while $\mathbf{w}^-(\mathbf{v})$ is repulsed. The strength of attraction and repulsion is obtained by the gradients of the cost function and the according learning rates. The trained model is a nearest neighbor classifier, i. e. an incoming data point is assigned to the same class as the nearest prototype. In general, the GLVQ is a sparse model with interpretative prototypes. The complexity of the model can be chosen by the user by specifying the number of prototypes per class. If only one prototype per class and the Euclidean distance is applied, GLVQ is a linear classifier. A more detailed description of the algorithm can be found in [44, 45], Fig. 3 provides a graphical representation.



		true class		
		EFR	**LFR**	total
predicted class	**EFR**	**True Positives (TP)**	**False Positives (FP)**	predicted EFR
	LFR	**False Negatives (FN)**	**True Negatives (TN)**	predicted LFR
		true EFR	true LFR	

$$\mathrm{PR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$

$$\mathrm{RE} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

$$\mathrm{CA} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$$

$$\mathrm{F}_1 = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}$$

Fig. 2 Confusion matrix. Exemplified CM with the formulas of precision (PR), recall (RE), accuracy (CA), and F_1-measure

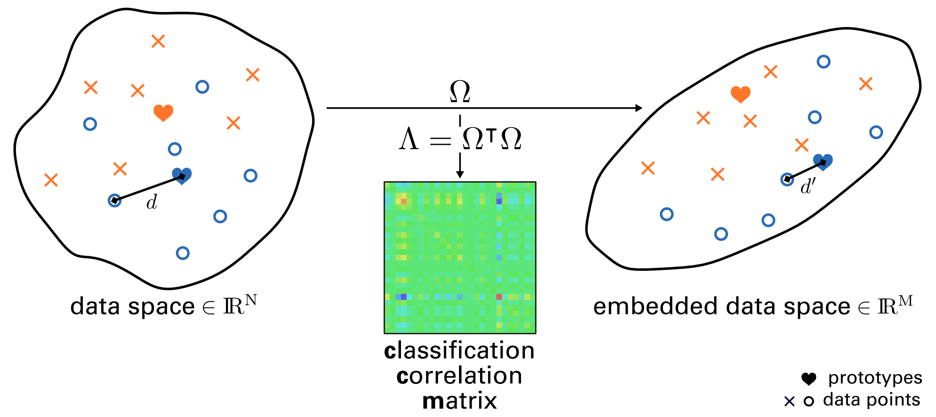

Ω
$\Lambda = \Omega^\intercal\Omega$
d
d'
data space $\in \mathbb{R}^N$
classification correlation matrix
embedded data space $\in \mathbb{R}^M$
prototypes
× ○ data points

Fig. 3 Principle of Generalized Matrix Learning Vector Quantization. Graphical depiction of learning with GMLVQ [63, 64]. One or multiple prototypes represent classes: each data point in the data space of dimension *N* belongs to the class of the prototype with the closest distance *d*. Prototypes are updated during learning as in LVQ [65]. Additionally, the matrix Ω maps the data space to an embedded data space of dimension *M*, where mapped distances d' are optimized. The matrix $\Lambda = \Omega'\Omega$ (CCM) represents the impact of each feature on the classification performance

A prominent extension of the GLVQ is the Matrix GLVQ [21]. Beside the prototypes, a mapping of the data points is learned for better separation of the classes (Fig. 4). This linear mapping, denoted by $\Omega \in \mathbb{R}^{M\times D}$, is powerful and provides additional information about the classification problem. Thereby, D is the number of features. The parameter m can be chosen by the user and indicates the mapping dimension. If the mapping dimension is equal to D, the matrix is quadratic, but M can also be set to values smaller than D, e. g. down to $M = 2$. In the latter case the GMLVQ can be used for visualization of the dataset by mapping the dataset into the two-dimensional space [46]. Moreover, the matrix $CCM = \Omega'\Omega$ is termed Classification Correlation Matrix (CCM) [44]. In contrast to the correlation matrix of the features, the CCM reflects the correlations between them under the aspect of class discrimination (Fig. 5b), i. e. positive or negative values of high magnitude between two features indicate a high positive or negative correlation of the features beneficial for the discrimination of classes. High values on the main diagonal occur for features important for the distinction of classes in general (see Fig. 5a).

Classification of Early Folding Residues

In the first step the dataset is standardized by z-score transformation. As mentioned before, the given dataset has a very unbalanced class distribution, i. e. only 482 data points of class EFR and 2784 of class LFR. In such cases the classification accuracy is inconclusive because it only takes correctly classified data points into account. Therefore, we determine further prominent evaluations measures based on the Confusion Matrix (CM) such as precision, recall, F_1-measure, and Area Under The Receiver Operating Characteristic (auROC) [47, 48]. The precision considers data points predicted as the positive class (here EFR) and recall on all data points, which are real positives. In our example, the number of EFR is drastically smaller than that of LFR, so in general the precision is much worse than recall. The F_1-measure, which is the harmonic mean between precision and recall, is sensitive if one of these values is getting too small. The Receiver Operating Characteristic

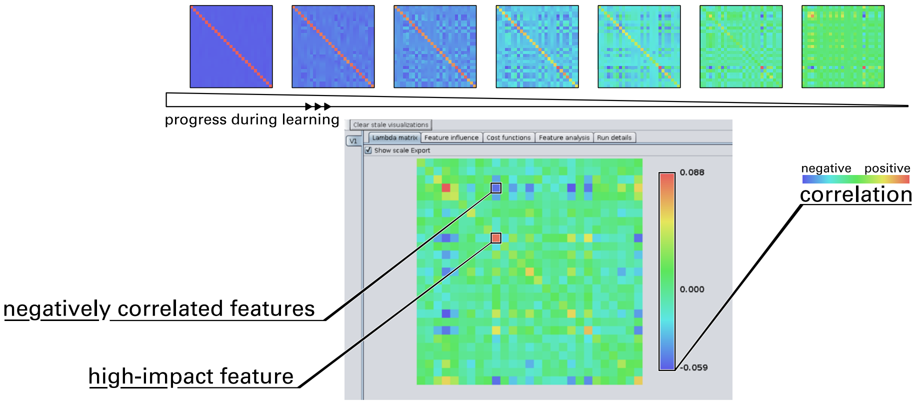


Fig. 4 The process of learning. The graphical user interface of the GMLVQ Weka implementation. The matrix panel shows the CCM and displays live updates during the learning process. A color bar represents the scale of the matrix elements with a coloring scheme similar to a heat map

(ROC) is a graphical plot illustrating the trade-off between true positives and false positives for a model. According to the Weka documentation, the ROC is obtained by varying the threshold on the class probability estimates.

We applied 10-fold cross validation on different classifiers to compare the results of GMLVQ to other state-of-the-art methods (see Table 2). Furthermore, we investigated different parameter settings of the GMLVQ in detail (to allow for an unbiased comparison, parameters for other methods were chosen by grid search to balance between performance and potential overfitting). On one side, the model size of the GMLVQ is a parameter chosen by the user. Here, we chose one prototype per class resulting in a linear classifier and five prototypes per class, which is more complex. Moreover, the GMLVQ has the feature to optimize other CM-based evaluation measures like the F_β-measure or a

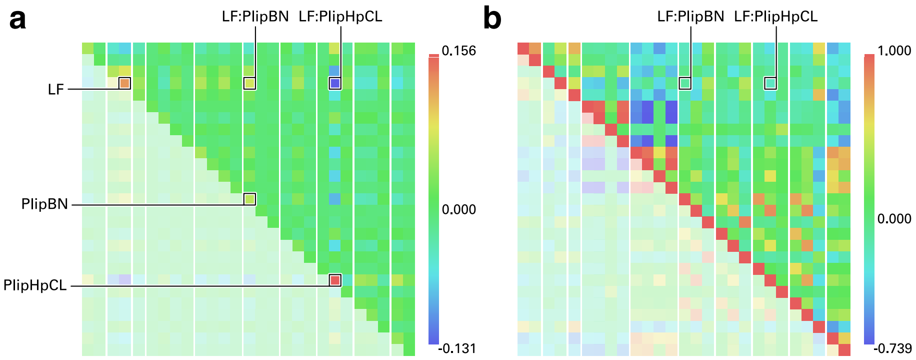


Fig. 5 The Classification Correlation Matrix for the classification of Early Folding Residues. **a** The CCM depicts the positive impact of individual features for the classification performance on its main diagonal. Especially, *PlipHpCL*, *LF*, and *PlipBN* are features which discriminate EFR and LFR. The influence of ordered secondary structure elements was shown before [17, 19]. Both betweenness and closeness centrality tend to be increased for EFR which indicates their importance for the assembly of secondary structure elements by long-range hydrophobic interactions [17]. Other entries of the matrix describe pairs of features which are positively (red) or negatively (blue) correlated and increase classification performance further. **b** The standard correlation matrix of all features of the whole dataset. Again, positive and negative correlations are depicted in red and blue respectively. Interestingly, the features pointed out by GMLVQ do not stand out. Vice versa, strong correlations between features do not imply a favorable influence on the classification performance

Table 2 CM presents the confusion matrix of a run. The first row captures the number of true positives and false positives. The second row presents the number of false negatives and true negatives. The test results in % right of CM and algorithmic parameters used for the classification of the data determined with Weka

CM		CA	PR	RE	F_1	auROC
Naive Bayes						
187	195	72.8	23.9	38.8	29.6	70.9
195	2190					
Random Forest						
192	290	82.1	39.6	39.8	39.7	64.7
293	2491					
Support Vector Machine						
134	348	**87.0**	**63.2**	27.8	38.6	62.5
78	2706					
GMLVQ with 1 prototype per class						
Run 1						
320	162	69.6	27.8	66.4	39.2	67.7
830	1954					
Run 2						
351	162	68.7	28.3	**72.8**	**40.7**	73.7
890	1954					
Run 3						
348	134	68.6	28.1	72.2	40.4	**76.6**
891	1893					
GMLVQ with 5 prototype per class						
Run 4						
187	295	77.4	29.7	38.8	33.6	69.4
443	2341					
Run 5						
288	194	69.0	26.0	59.8	36.2	70.5
819	1965					
Run 6						
274	208	70.3	26.4	56.8	36.1	70.3
763	2021					

Additionally, we marked the best values for the single evaluation measured bold. If not stated otherwise, default setup was used. SVM with RBF-kernel ($\sigma = 5$) which results in 1193 number of support vectors. Weights for weighted accuracy: 0.75 and 0.25. F_β-measure with $\beta = 1$ (F_1)

linear combination of precision and recall. These can take the unbalanced class distribution into account. These aspects are reflected in Table 2. The comparison of the different classification models is challenging. It is difficult to decide objectively which classifier performs best. The SVM ends up with the best accuracy, yet the recall is very low. On the other side, the GMLVQ optimizing the weighted accuracy has the best recall and F_1-value and optimizing the F_β-measure ends up with the best value in the auROC. Furthermore, we can notice that very complex models do not automatically perform better. The Naive Bayes (NB), a very simple, fast and linear classifier performs comparable to the other much more complex models like Random Forest (RF) or SVM, which utilizes 1193 support vectors, i. e. 36% of the data points are necessary to describe the hyperplane. The GMLVQ runs with five prototypes per class perform better in training than GMLVQ with one prototype, yet, in test the sparse model is more suitable. We applied different cost functions evaluating approximated values of classification accuracy, weighted classification accuracy, F_1-measure, or weighted precision-recall. The results with the according parameter selection (Table 3) are listed in Table 2.

To sum up, GMLVQ provides better results in recall even if the model is chosen to be very sparse. Distinguishing EFR and LFR is challenging and a clear separation was not achievable using the described features. GMLVQ was trained on the dataset in order to retrieve the most discriminative features of EFR and to showcase the capabilities and handling of the visualization.

Visualization of learning process and interpretation of classification results

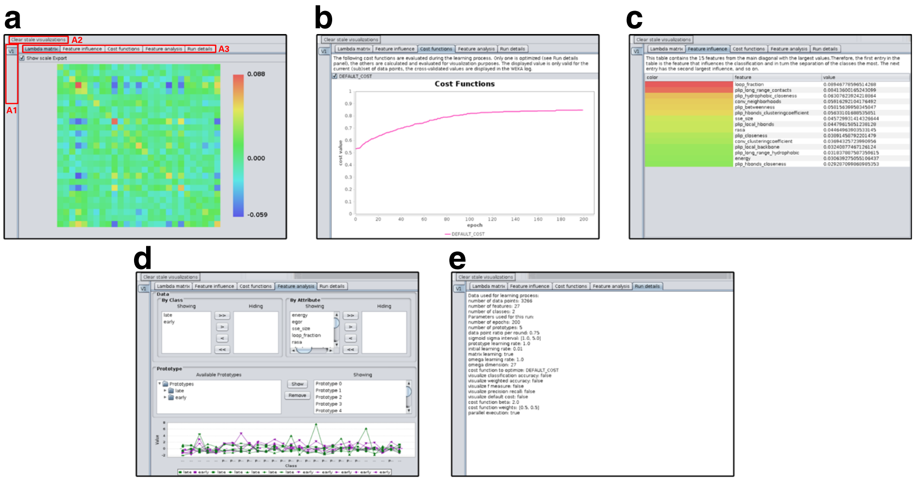
The GMLVQ plug-in tracks and summarizes each run by various visualization panels (Fig. 6): the CCM panel (Fig. 6a), the cost function panel (Fig. 6b), the feature influence panel (Fig. 6c), the feature analysis panel which depicts the prototype placement (Fig. 6d), and the run details panel which reports the parameters of the corresponding run (Fig. 6e). A detailed description on the example for the EFR dataset is given in order to demonstrate how results of GMLVQ can be interpreted by integrating information of these visualization panels.

For the presented dataset, the CCM (Fig. 5a) is primarily homogeneous which is indicated by values close to zero. The major contributing features are the *LF*, *PlipBN*, and especially *PlipHpCL* as these features exhibit the highest scores on the main diagonal of the CCM. The positive correlation of *LF* and *PlipBN* contributes to the classification performance as indicated by positive values described by the corresponding element. Also, the negative correlation of *PlipHpCL* to both features increases classification performance. The *PlipHpCL* is negatively correlated to various other features such as *SecSize*, *PlipLR*, *PlipHbLR*, and *PlipHbCL*. To a lesser degree, *e* and *PlipNC* are associated positively. It has to be pointed out that the CCM differs substantially from the correlation matrix (see Fig. 5b). In the correlation matrix, strong positive correlations are present in the fourth group of features (local contact counts) and negative correlations in the fifth group (long-range contact counts). Relevant associations between features pointed out by GMLVQ are not obvious from the correlation matrix. The five most important features for discrimination are listed in Table 4 which was derived from the feature influence panel (Fig. 6c). The prototype placement depicted in the feature analysis panel (Fig. 6d)

Table 3 Parameter selection to obtain the results of Table 2 using the Weka plug-in

Parameter	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6
Cost function to optimize	CA	WCA	F_1	CA	WCA	F_1
Number of epochs	150	150	150	250	250	250
Number of prototypes	1	1	1	5	5	5
Data point ratio per round	0.75	0.75	0.75	0.75	0.75	0.75
Sigmoid sigma interval	[1.0,5.0]	[1.0,15.0]	[1.0,50.0]	[1.0,5.0]	[1.0,15.0]	[1.0,50.0]
Prototype learning rate	1.0	1.0	1.0	1.0	1.0	1.0
Matrix learning	True	True	True	True	True	True
Omega learning rate	1.0	1.0	1.0	1.0	1.0	1.0
Omega dimension	27	27	27	27	27	27
Cost function beta	-	-	1	-	-	1
Cost function weights	-	[0.75,0.25]	-	-	[0.75,0.25]	-
Parallel execution	True	True	True	True	True	True

Classification accuracy (CA), weighted classification accuracy (WCA) with weights 0.75 and 0.25 as well as F_β-measure with $\beta = 1$ (F_1)



The following cost functions are evaluated during the learning process. Only one is optimized (see Run details panel), the others are calculated and evaluated for visualization purposes. The displayed value is only valid for the current (sub)set of data points, the cross-validated values are displayed in the WEKA log.

This table contains the 15 features from the main diagonal with the largest values.Therefore, the first entry in the table is the feature that influences the classification and in turn the separation of the classes the most. The next entry has the second largest influence, and so on.

color	feature	value
🟥	loop_fraction	0.08946778596514268
🟥	plip_long_range_contacts	0.08413600165243099
🟧	plip_hydrophobic_closeness	0.06307623924218864
🟧	conv_neighborhoods	0.05916292104176492
🟨	plip_betweenness	0.05815639958345847
🟨	plip_hbonds_clusteringcoefficient	0.05633101688535851
🟨	sse_size	0.04572993141432664
🟨	plip_local_hbonds	0.04479615851238128
🟨	rasa	0.04464963903533145
🟩	plip_closeness	0.03891458792201479
🟩	conv_clusteringcoefficient	0.03694325723990956
🟩	plip_local_backbone	0.03240877467126124
🟩	plip_long_range_hydrophobic	0.03183788758735961
🟩	energy	0.03063927505510643
🟩	plip_hbonds_closeness	0.02928709986898535

Data used for learning process:
number of data points: 3266
number of features: 27
number of classes: 2
Parameters used for this run:
number of epochs: 200
number of prototypes: 5
data point ratio per round: 0.75
sigmoid sigma interval: [1.0, 5.0]
prototype learning rate: 1.0
initial learning rate: 0.01
matrix learning: true
omega learning rate: 1.0
omega dimension: 27
cost function to optimize: DEFAULT_COST
visualize classification accuracy: false
visualize weighted accuracy: false
visualize f measure: false
visualize precision recall: false
visualize default cost: false
cost function beta: 2.0
cost function weights: [0.5, 0.5]
parallel execution: true

Fig. 6 Screenshots of the Weka plug-in for GMLVQ. **a** The visualization of the CCM. The color scale indicates positive or negative correlations. (**a1**) The visualizations of each separate run will appear in this area. By clicking on the respective tab, one can easily switch between individual runs, e.g. cross validation runs. (**a2**) This button clears all visualizations except the latest. (**a3**) The tabs allow the user to switch between different visualizations of the current run. **b** The chart visualizes the cost functions over the course of learning. Additional functions can be visualized here, alongside with the cost function which is optimized. **c** The feature influence of single features of the current run. The top-ranked features have the highest contribution for the classification performance. **d** The feature analysis panel allows the detailed investigation of features and prototypes. **e** This panel shows the parameters which were used for the current run

describes which values individual features adapt for optimal classification performance. This information is not evident from the CCM but necessary for the interpretation of the learned model. Selecting only these five features and learning a model on this dimensionality-reduced dataset, shows a performance similar to the full model. GMLVQ with weighted accuracy and one prototype per class is given in Table 5. Recall and F_1 value are even better compared to using all features. Thus, the GMLVQ can also be used for feature extraction.

The homogeneity observed in the CCM is the result of the similarity of several features. At a trivial level, topological descriptors computed on differing graph definitions are likely to result in redundant information. In that case, it is coincidental which feature will be highlighted even though all other correlated features capture similar information. Even if

Table 4 Summary of the top five features which are most important for the classification of EFR according to the GMLVQ and RF method

	GMLVQ		Random Forest	
Feature	Rank	Influence score	Rank	Influence score
PlipHpCL	1	0.159	3	1.370
LF	2	0.127	2	1.403
PlipBN	3	0.063	15	0.900
SecSize	4	0.059	19	0.854
e	5	0.042	4	1.332
PlipCL	7	0.012	1	1.700
ConvCC	23	0.009	5	1.223

Importance scores for the RF were computed by the MATLAB implementation. Influence scores are in arbitrary units, higher values occur for features important for class discrimination. The values of GMLVQ and the predictor importance values are method-specific and not directly comparable; therefore, the ranks of the top five features are given

Table 5 CM presents the confusion matrix of a run. The first row captures the number of true positives and false positives. The second row presents the number of false negatives and true negatives. Performance of GMLVQ using only the five most important features

CM		CA	PR	RE	F_1	auROC
376	106	67.7	28.4	78.0	41.6	69.0
950	1834					

such features are strongly correlated, the CCM will only capture these characteristics if the correlation also contributes to the classification performance.

The *PlipBN* feature is the betweenness centrality [41, 42] derived from all contacts such as hydrogen bonds or hydrophobic interactions [39] in a protein structure. For this graph, residues with many of the shortest paths passing through them exhibit high betweenness centrality scores. This feature is highly discriminative for EFR and LFR as captured in the CCM. The prototypes which represent the EFR class display above average *PlipBN* values, indicating that EFR are better connected in the residue graph than their LFR counterparts. In fact, EFR exhibit a higher degree and are crucial connectors, so-called hubs. Residues with high betweenness centrality values have been shown to be crucial for the formation of stable, local structure and often constitute the folding nucleus of proteins [4, 42, 49].

The *LF* is relatively low for EFR which implies that EFR tend to be surrounded by ordered secondary structure elements. Analogously, this is negatively correlated to the size of the surrounding secondary structure elements and positively correlated to the *Rasa* values as it has been shown in previous studies [11, 12, 19, 42]. The *LF* feature is furthermore negatively correlated to *e* which indicates that ordered secondary structure elements result in favorable, low energy local conformations. These local structures are assumed to form autonomously and guide the folding process [12, 18].

The importance of the *PlipHpCL* represents the relevance of hydrophobic interactions in the core of protein structures (Fig. 7). EFR have an increased propensity to occur in the core of protein structures which is isolated from the polar solvent [8, 19]. However, a buried or exposed state [50] derived from the *Rasa* feature cannot explain the origin and characteristics of EFR [17]. The closeness centrality [51] is defined as the inverse of the average path length of a residue to all other residues in the graph. It describes how well connected individual residues are which is a similar characteristic as covered by the betweenness centrality [41, 42]. The fact that both *PlipBN* and *PlipHpCL* are the most influential features for the classification demonstrates that they still capture slightly different aspects. The classification performance benefits from a negative correlation of both features. EFR occur primarily in the hydrophobic core of a structure where they participate in an increased number of hydrophobic interactions with surrounding residues. Previously, hydrophobic interactions have been shown to be relevant for the initiation and guidance of the protein folding process itself as well as its in silico modeling [2, 52–54]. They can be realized by a subset of amino acids and have an increased propensity to form ordered regions [11, 32]. The importance of the *PlipHpCL* feature and the placement of the prototypes implies that EFR are primarily embedded in the hydrophobic network of protein structures. EFR have been previously described to form more hydrophobic interactions which are important for the correct assembly of protein regions separated at sequence level [17].

In summary, the visualized classification of the GMLVQ run pointed out that many features capture redundant information. A subset of the features (*PlipHpCL*, *LF*, and *PlipBN*)

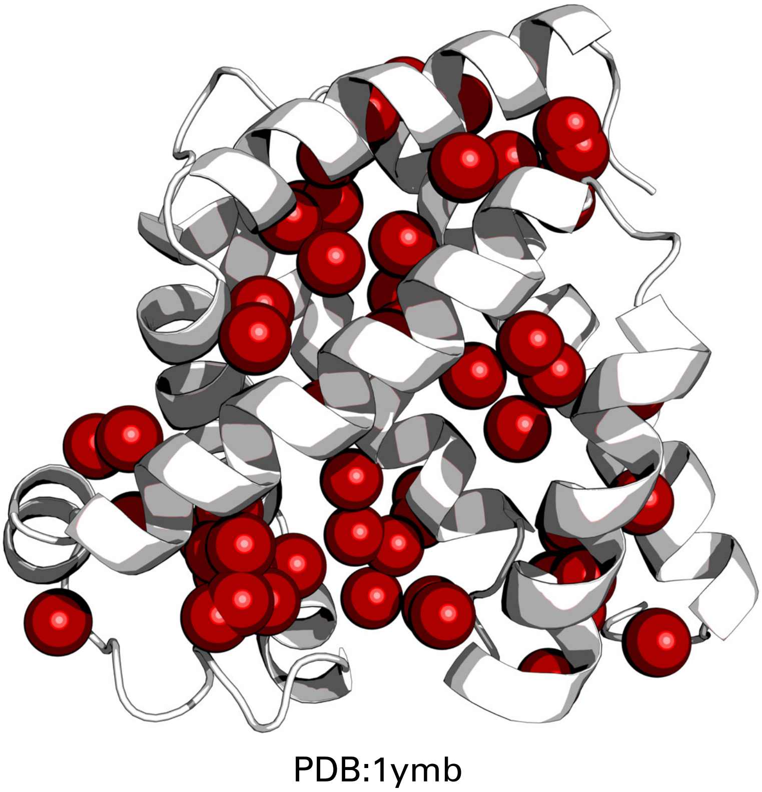


Fig. 7 Rendering of the network of hydrophobic interactions. Structure of horse heart myoglobin (PDB:1ymb). In this structure, 58 hydrophobic interactions were detected by PLIP [39]. The centroids between interacting residues are depicted as red spheres. This highlights the strong contribution of hydrophobic interactions in the protein core

is discriminative for both classes. Their importance and their respective correlations are in agreement with previous studies on EFR [16, 19] and, more general, folding nuclei [8, 12, 18, 42, 55].

Other methods such as RF are also capable of reporting the most influential features for a classification problem. The reported scores of GMLVQ and RF were ranked to make them comparable (Table 4). Some features such as *LF* or *e* are identified as high influence features independently of classification method. These features capture unique properties, which are not described by other features. In contrast the topological properties (*PlipHpCL*, *PlipBN*, *PlipCL*, and *ConvCC*) tend to describe similar properties and constitute redundant features. Their reported influence deviates heavily between GMLVQ and RF. It is remarkable that the most influential feature in either case is the closeness centrality. According to GMLVQ it is the value computed using the network of hydrophobic interactions (*PlipHpCL*) whereas RF identifies the closeness centrality computed using all non-covalent interactions (*PlipCL*) as most influential. RF ranks *PlipHpCL* as the third most influential feature which substantiates the importance of the previously discussed network of hydrophobic interactions (Fig. 7).

Conclusion

Machine as well as deep learning are trending in (life) sciences. Yet, a lot of classification problems are difficult to solve. Especially for problems with highly unbalanced class distributions the choice of the best model is crucial. Beside evaluation measures, other properties might be essential to select a suitable classifier. One key aspect is the interpretability of the learning process and the resulting model. GMLVQ is a prototype-based classifier. GMLVQ provides an interpretable classification model and was integrated into the Weka framework to make this classifier and its visualization capabilities accessible to a wide range of scientists.

A dataset of key residues of the protein folding process was investigated. GMLVQ performs comparable to other state-of-the-art methods such as SVM or RF but provides a readily interpretable classification model. From a set of 27 features, GMLVQ identified the fraction of ordered secondary structure elements, the betweenness centrality based on non-covalent contacts, and the closeness centrality using only hydrophobic interactions as the most relevant features for the distinction between Early and Late Folding Residues. Despite the specific use case on protein folding, the GMLVQ classifier is generally applicable for classification problems and constitutes a valuable addition to toolkit of bioinformatics [56–61].

The classification performance may be improved by using additional features; however, for sake of simplicity such features were omitted because their computation would require additional algorithms or models. Promising candidates are backbone rigidity values [11], sequence-based predictions of Early Folding Residues [16], or evolutionary coupling scores [62]. All of them have been previously shown to be discriminative for Early Folding Residues [16, 19] and may increase the classification performance of this exemplary application of the Weka plug-in. Established sequence-based features employed in other classification models [27, 28] may further enhance the prediction of Early Folding Residues.

Additional files

Additional file 1: Dataset in ARFF. (ARFF 577 kb)

Additional file 2: Dataset as CSV. (CSV 576 kb)

Additional file 3: Correlation matrix of all features. (CSV 13.2 kb)

Acknowledgments
Not applicable.

Funding
This work was funded by the European Social Fund (http://ec.europa.eu/esf/, grant numbers: 100235472 for FK and 1463643413126 for CL), the Free State of Saxony and the Saxon Ministry of Fine Arts. The funding agencies had no impact on design or conduct of the study.

Availability of data and materials
All data generated or analyzed during this study is included in this published article and its supplementary information files. An open-source release and a step-by-step guide for the installation as well as usage of the GMLVQ Weka plug-in is available at https://github.com/JonStargaryen/gmlvq. The current version of the plug-in is deposited at https://doi.org/10.5281/zenodo.1326272. A web server employing the trained classification model is accessible at https://biosciences.hs-mittweida.de/efpred/.

Authors' contributions
MK and TV designed the GMLVQ algorithm which was implemented in Java by SB, CL, and FK. SB and MK analyzed the data. SB, MK, CL, and FK drafted the manuscript. TV and DL supervised the project. All authors read and approved the 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

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 6 August 2018 Accepted: 20 November 2018
Published online: 05 January 2019

References

1. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19(1):15.
2. Dill KA, Ozkan SB, Shell MS, Weikl TR. The protein folding problem. Annu Rev Biophys. 2008;37:289–316.
3. Haglund E, Danielsson J, Kadhirvel S, Lindberg MO, Logan DT, Oliveberg M. Trimming down a protein structure to its bare foldons: spatial organization of the cooperative unit. J Biol Chem. 2012;287(4):2731–8.
4. Vendruscolo M, Paci E, Dobson CM, Karplus M. Three key residues form a critical contact network in a protein folding transition state. Nature. 2001;409(6820):641–5.
5. Dokholyan NV, Li L, Ding F, Shakhnovich EI. Topological determinants of protein folding. Proc Natl Acad Sci. 2002;99(13):8637–41.
6. Roder H, Elove GA, Englander SW. Structural characterization of folding intermediates in cytochrome c by H-exchange labelling and proton NMR. Nature. 1988;335(6192):700–4.
7. Bai Y, Sosnick TR, Mayne L, Englander SW. Science. 1995;269(5221):192–7.
8. Li R, Woodward C. The hydrogen exchange core and protein folding. Protein Sci. 1999;8(8):1571–90.
9. Chu R, Pei W, Takei J, Bai Y. Relationship between the native-state hydrogen exchange and folding pathways of a four-helix bundle protein. Biochemistry. 2002;41(25):7998–8003.
10. Englander SW, Mayne L, Krishna MM. Protein folding and misfolding: mechanism and principles. Q Rev Biophys. 2007;40(4):287–326.
11. Cilia E, Pancsa R, Tompa P, Lenaerts T, Vranken WF. From protein sequence to dynamics and disorder with DynaMine. Nat Commun. 2013;4:2741.
12. Englander SW, Mayne L. The nature of protein folding pathways. Proc Natl Acad Sci. 2014;111(45):15873–80.
13. Pancsa R, Varadi M, Tompa P, Vranken WF. Start2fold: a database of hydrogen/deuterium exchange data on protein folding and stability. Nucleic Acids Res. 2016;44(D1):429–34.
14. Baldwin RL, Rose GD. Is protein folding hierarchic? i. local structure and peptide folding. Trends Biochem Sci. 1999;24(1):26–33.
15. Baldwin RL, Rose GD. Is protein folding hierarchic? ii. folding intermediates and transition states. Trends Biochem Sci. 1999;24(2):77–83.
16. Raimondi D, Orlando G, Pancsa R, Khan T, Vranken WF. Exploring the sequence-based prediction of folding initiation sites in proteins. Sci Rep. 2017;7(1):8826.
17. Bittrich S, Schroeder M, Labudde D. Characterizing the relation of functional and early folding residues in protein structures using the example of aminoacyl-trna synthetases. PLoS ONE. 2018;13(10):1–23.
18. Englander SW, Mayne L. The case for defined protein folding pathways. Proc Natl Acad Sci. 2017;114(31):8253–8.
19. Pancsa R, Raimondi D, Cilia E, Vranken WF. Early folding events, local interactions, and conservation of protein backbone rigidity. Biophys J. 2016;110(3):572–83.
20. Sato A, Yamada K. Generalized learning vector quantization. In: Touretzky DS, Mozer MC, Hasselmo ME, editors. Advances in Neural Information Processing Systems 8. Cambridge: MIT Press; 1996. p. 423–9.
21. Schneider P, Biehl M, Hammer B. Distance learning in discriminative vector quantization. Neural Comput. 2009;21(10):2942–69.
22. Holmes G, Donkin A, Witten IH. Weka: A machine learning workbench. In: Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference On. New York: IEEE; 1994. p. 357–61.
23. Frank E, Hall M, Trigg L, Holmes G, Witten IH. Data mining in bioinformatics using weka. Bioinformatics. 2004;20(15):2479–81.
24. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The weka data mining software: an update. ACM SIGKDD Explor Newsl. 2009;11(1):10–18.
25. Wei L, Luan S, Nagai LAE, Su R, Zou Q. Exploring sequence-based features for the improved prediction of dna n4-methylcytosine sites in multiple species. Bioinformatics. 2018;824. [Epub ahead of print].
26. Wei L, Chen H, Su R. M6apred-el: A sequence-based predictor for identifying n6-methyladenosine sites using ensemble learning. Mol Therapy-Nucleic Acids. 2018;12:635–44.
27. Wei L, Xing P, Shi G, Ji Z-L, Zou Q. Fast prediction of protein methylation sites using a sequence-based feature selection technique. IEEE/ACM Trans Comput Biol Bioinform. 2017;1:1–1.
28. Wei L, Zhou C, Chen H, Song J, Su R. Acpred-fl: a sequence-based predictor based on effective feature representation to improve the prediction of anti-cancer peptides. Bioinformatics. 2018;34(23):4007–4016.
29. Rose AS, Hildebrand PW. Nucleic Acids Res. 2015;43(W1):576–9.
30. Rose AS, Bradley AR, Valasatava Y, Duarte JM, Prlić A, Rose PW. Web-based molecular graphics for large complexes. In: Proceedings of the 21st International Conference on Web3D Technology. New York: ACM; 2016. p. 185–6.
31. Anfinsen CB, Scheraga HA. Experimental and theoretical aspects of protein folding. Adv Protein Chem. 1975;29: 205–300.
32. Heinke F, Schildbach S, Stockmann D, Labudde D. epros—a database and toolbox for investigating protein sequence–structure–function relationships through energy profiles. Nucleic Acids Res. 2012;41(D1):320–6.
33. Bittrich S, Heinke F, Labudde D. Beyond Databases, Architectures and Structures. Advanced Technologies for Data Mining and Knowledge Discovery. BDAS 2015, BDAS 2016. Communications in Computer and Information Science, vol 613. Cham: Springer; 2016. pp. 419–33.
34. Kabsch W, Sander C. Dssp: definition of secondary structure of proteins given a set of 3d coordinates. Biopolymers. 1983;22:2577–637.

35. Holland RC, Down TA, Pocock M, Prlić A, Huen D, James K, Foisy S, Dräger A, Yates A, Heuer M, et al. Biojava: an open-source framework for bioinformatics. Bioinformatics. 2008;24(18):2096–7.
36. Prlić A, Yates A, Bliven SE, Rose PW, Jacobsen J, Troshin PV, Chapman M, Gao J, Koh CH, Foisy S, et al. Biojava: an open-source framework for bioinformatics in 2012. Bioinformatics. 2012;28(20):2693–5.
37. Benkert P, Künzli M, Schwede T. Qmean server for protein model quality estimation. Nucleic Acids Res. 2009;37(suppl_2):510–4.
38. Shrake A, Rupley J. Environment and exposure to solvent of protein atoms. lysozyme and insulin. J Mol Biol. 1973;79(2):351–71.
39. Salentin S, Schreiber S, Haupt VJ, Adasme MF, Schroeder M. Plip: fully automated protein–ligand interaction profiler. Nucleic Acids Res. 2015;43(W1):443–7.
40. Adhikari B, Cheng J. Improved protein structure reconstruction using secondary structures, contacts at higher distance thresholds, and non-contacts. BMC Bioinformatics. 2017;18(1):380.
41. Freeman LC. A set of measures of centrality based on betweenness. Sociometry. 1977;40(1):35–41.
42. Vendruscolo M, Dokholyan NV, Paci E, Karplus M. Small-world view of the amino acids that play a key role in protein folding. Phys Rev E. 2002;65(6):061910.
43. Kohonen T. Learning vector quantization for pattern recognition. Technical report, TKK-F-A601, Helsinki. 1986.
44. Kaden M, Lange M, Nebel D, Riedel M, Geweniger T, Villmann T. Aspects in classification learning-review of recent developments in learning vector quantization. Found Comput Dec Sci. 2014;39(2):79–105.
45. Kaden M. Integration of auxiliary data knowledge in prototype based vector quantization and classification models. PhD thesis, University Leipzig, Germany. 2015.
46. Bunte K, Schneider P, Hammer B, Schleif F, Villmann T, Biehl M. Limited rank matrix learning, discriminative dimension reduction and visualization. Neural Netw. 2012;26:159–73.
47. Chawla NV. Data Mining for Imbalanced Datasets: An Overview. In: Maimon O, Rokach L, editors. Data Mining and Knowledge Discovery Handbook. Boston: Springer; 2010. p. 875–86.
48. Fawcett T. An introduction to roc analysis. Pattern Recogn Lett. 2006;27(8):861–74.
49. Brinda K, Vishveshwara S. A network representation of protein structures: implications for protein stability. Biophys J. 2005;89(6):4159–70.
50. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Protein Struct Funct Bioinform. 1994;20(3):216–26.
51. Amitai G, Shemesh A, Sitbon E, Shklar M, Netanely D, Venger I, Pietrokovski S. Network analysis of protein structures identifies functional residues. J Mol Biol. 2004;344(4):1135–46.
52. Dill KA. Theory for the folding and stability of globular proteins. Biochemistry. 1985;24(6):1501–9.
53. Faísca PF. The nucleation mechanism of protein folding: a survey of computer simulation studies. J Phys Condens Matter. 2009;21(37):373102.
54. Gromiha MM. Multiple contact network is a key determinant to protein folding rates. J Chem Inf Model. 2009;49(4): 1130–5.
55. Mirny LA, Shakhnovich EI. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function1. J Mol Biol. 1999;291(1):177–96.
56. Alegre E, Biehl M, Petkov N, Sánchez L. Automatic classification of the acrosome status of boar spermatozoa using digital image processing and lvq. Comput Biol Med. 2008;38(4):461–8.
57. Biehl M, Schneider P, Smith D, Stiekema H, Taylor A, Hughes B, Shackleton C, Stewart P, Arlt W. Matrix relevance lvq in steroid metabolomics based classification of adrenal tumors. In: ESANN 2012 proceedings, European Symposium on Artificial Neural Networks. Bruges: Computational Intelligence and Machine Learning; 2012.
58. Schneider P, Schleif F-M, Villmann T, Biehl M. Generalized matrix learning vector quantizer for the analysis of spectral data. In: ESANN 2008 proceedings, European Symposium on Artificial Neural Networks. Bruges: Computational Intelligence and Machine Learning; 2008.
59. Kästner M, Strickert M, Villmann T, Mittweida S-G. A sparse kernelized matrix learning vector quantization model for human activity recognition. In: ESANN 2013 proceedings, European Symposium on Artificial Neural Networks. Bruges: Computational Intelligence and Machine Learning; 2013.
60. Prahm C, Paassen B, Schulz A, Hammer B, Aszmann O. Transfer learning for rapid re-calibration of a myoelectric prosthesis after electrode shift. In: Converging Clinical and Engineering Research on Neurorehabilitation II. Cham: Springer; 2017. p. 153–7.
61. Mudali D, Biehl M, Leenders KL, Roerdink JB. Lvq and svm classification of fdg-pet brain data. In: Advances in Self-Organizing Maps and Learning Vector Quantization. Cham: Springer; 2016. p. 205–15.
62. Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation. Nat Biotechnol. 2012;30(11): 1072–80.
63. Hammer B, Villmann T. Generalized relevance learning vector quantization. Neural Netw. 2002;15(8):1059–68.
64. Kästner M, Hammer B, Biehl M, Villmann T. Functional relevance learning in generalized learning vector quantization. Neurocomputing. 2012;90:85–95.
65. Kohonen T. Learning vector quantization. In: Self-Organizing Maps. Berlin: Springer; 1997. p. 203–17.