An Evaluation of Feature Selection Approaches in Finding Amyloidogenic Regions in Protein Sequences

Smitha Sunil Kumaran Nair
Department of Computer Science and Engineering
Manipal Institute of Technology
Manipal University, Karnataka, India

N. V. Subba Reddy
Mody Institute of Technology and Science University
Rajasthan
India

Hareesha K. S
Department of Computer Science and Engineering
Manipal Institute of Technology
Manipal University, Karnataka, India

ABSTRACT
Amyloidogenic regions in polypeptide chains are associated with a number of diseases. Experimental evidence is compelling in favor of the hypothesis that small segments of proteins are responsible for its amyloidogenic behavior. Thus, identifying these short peptides is critical for understanding diseases associated with protein misfolding and developing sequence-targeted anti-aggregation drugs. The *in silico* approaches using phenomenological models based on bio-physio-chemical properties of amino acids suffer from "curse of dimensionality". Therefore, before adopting standard classification algorithms to predict such fibril motifs, the "curse of dimensionality" needs to be solved. The present study evaluates the performance of feature selection algorithms namely filter, wrapper and embedded models in conjunction with Support Vector Machine classifier. We also propose a novel integrated feature selection strategy based on Genetic Algorithm and Support Vector Machine to get an optimal number of features in predicting the amyloid fibril-forming short stretches of peptides. In addition, we investigated the performances of feature selection models that resulted in new and complementary set of properties and concludes that the proposed integrated dimensionality reduction technique outperforms all other methods and achieves the highest sensitivity and specificity of 86% and 82% respectively.

General Terms
Bioinformatics, Proteomics, Feature selection

Keywords
Amyloid fibril, physicochemical properties, Genetic Algorithm, Support Vector Machine

1. INTRODUCTION
Amyloid fibril formation is widely observed in human pathologies such as, Alzheimer's disease, Parkinson's disease, Huntington's disease and Type II diabetes. In these diseases, proteins with unrelated sequences aggregate to form highly characteristic amyloid fibrils. There is currently no effective treatment against these progressive disorders, most of which affect the brain in a devastating way. Therefore, it is of fundamental medical interest to understand the mechanisms of fibrillogenesis with the ultimate goal of determining the mature fibrils [1].

Recent studies have proved that not all regions of a polypeptide are equally important for determining its aggregation tendency, but very short, continuous and specific amino acid stretches that would act as facilitators or inhibitors of amyloid fibril formation [2]. The knowledge of such short peptide sequences and their location are important for the development of targeted strategies to combat diseases associated with amyloid formation [3] and also help in understanding the mechanism of amyloid formation that leads to effective treatments for amyloid illnesses [4].

Recent efforts in understanding the physicochemical grounds [3] and structural denominators [1] of amyloid fibril formation has led to the development of several algorithms, capable of predicting a number of aggregation related parameters of a protein directly from its amino acid sequence. The determination of the physicochemical principles underlying amyloid deposition is fundamental to the identification of therapeutic strategies to prevent or cure amyloid-related disorders. Bioinformatics tools that perform prediction tasks are increasingly incorporating physico-chemical property based metric to increase their performance and to derive knowledge based rules [28].

As research continues for the understanding of the mechanisms involved in amyloid formation, the development of prediction methods is an important complement to experimental approaches [20]. From a biological perspective, the significance of features and their values in identifying potential biomarkers using supervised training of classifiers should be investigated [34]. A prerequisite for this task is to design an efficient and effective feature selection model. In this article, we propose a novel integrated feature selection scheme based on Genetic Algorithm (GA) [16] and Support Vector Machine (SVM) that purely follows a sequence-based design strategy to select significant bio-physio-chemical properties of amino acids from Amino Acid index database in DBGet (Japan) and ProtScale in Swiss Expasy to represent protein sequence features thereby reducing the dimensionality of the input space that would improve the overall classification performance in predicting the amyloid motifs in proteins. The effectiveness of several feature selection techniques considered in the present study is evaluated using SVM classifier on a collective dataset mentioned in the following section.

2. DATA AND METHODS

2.1 Biologically relevant sequence dataset

We compiled experimentally proved proteins related to amyloidosis and proteins with no experimentally determined amyloidogenic regions published in literature [2], [5], [6], [8], [20], [21], [26], [27], [33], in order to construct the training and testing dataset. All protein sequences were downloaded in Fasta format from Uniprot-Swissprot database [9]. The experimental analysis of different proteins which form amyloid fibrils revealed that these proteins contain rather small fragments which are required for the amyloidogenesis [31]. Michael J. Thompson et al., [8] claims that hexpeptides are sufficient for forming amyloid-like fibrils. Therefore, a dataset of hexpeptides including positive and negative examples of fibril formation was prepared by sliding a window of six residues. A set of 2512 hexmers of which 1232 that have been shown to form fibrils and 1280 that have yielded negative results in fibril-forming assays constitute the training data.

2.2 Feature mining

The overall capability of classifiers to predict fibril aggregates is based on the features used to encode the protein sequences. Since SVM classifier requires each data instance to be represented as a vector of real numbers [7], the numerical values of physicochemical or biochemical properties of amino acids are used to form the feature vector. The Amino Acid index (AAindex Version 9) [10] is a database that provides 544 properties associated with each of the 20 amino acids. Of the 544 indices, 13 have incomplete data, and were never considered.

According to Mathura et al., [28] properties that have missing values for any of the twenty amino acids and those that are less relevant to the study of protein sequence, structure and function are excluded in their database named APDbase [30]. Therefore, among all 531 features in [10], only 216 in APDbase were taken into account for the design of prediction algorithm.

Of the 246 entries in APDbase, the last 30 entries correspond to ProtScale in Swiss Expasy [29] which is not endowed with IDs or Accession Nos and the remaining 216 properties are from AAindex database [10]. The authors have designated certain Accession Nos in a similar fashion as those of in AAindex version 9 for the very last 30 properties available in [29]. Thus 246 indices were evaluated for potential use.

The values of each property were scaled so as to fall within a small specified range using z-score normalization (zero-mean normalization) technique [11], for it to be used by the classifier, and is mentioned below.

$$\nu' = \frac{\nu - \overline{A}}{\sigma_A} \quad (1)$$

This transforms the values of an attribute, A based on the mean and standard deviation of A . A value, ν of A is normalized to ν' by computing the equation (1) where $\overline{A}$ and σ_A are the mean and standard deviation, respectively, of attribute A.

2.3 Dimensionality reduction

To achieve a considerably better performance in terms of classification ability, it is a prerequisite to generate relevant features so as to discriminate well among classes. One of the most fundamental problems in bioinformatics, and machine learning is how to select a small subset of significant features. Literatures on the subject of Feature Selection (FS) are abundant, proposing taxonomy of FS algorithms [14], and presenting comparative studies. Optimal feature selection needs an exhaustive search through the space of feature subsets, and it is intractable when the number of features is large. For practical supervised learning algorithms, efficient methods are required [36].

As reviewed [14], FS methods can broadly fall into Filter model, Wrapper model and Embedded model. Filter methods assess the relevance of features by calculating feature relevance score not involving any learning algorithm, and low-scoring features are removed. They are computationally fast and simple, however the dependence among features are ignored. Wrapper methods apply a specific machine learning algorithm and utilize the corresponding classification performance to guide the feature selection. They have a high probability of producing classifiers with better classification performances than the filter approaches; but are very computationally intensive. In embedded techniques, the search for an optimal subset of features is built into the classifier construction, thus specific to a given learning algorithm. These methods have the advantage that they include the interaction with the classification model, while at the same time being far less computationally intensive than wrapper methods [14].

To exploit the advantages of filter, wrapper and embedded methods, we present integrated methods for subset selection. Initially a feature pre-selection is designed using filter as well as embedded approaches independently to exclude the irrelevant features so that the wrapper based method can derive the suitable features more efficiently, without searching through the whole feature space. The proposed two-stage routine for feature selection is discussed in the following section.

2.3.1 Feature pre-selection

In the present study, we did not start with a randomly chosen set of physicochemical properties which had been proved to be related to protein aggregation. In fact, we firstly evaluated every property in APDbase and selected an initial subset of properties using embedded model and filter based models as discussed below.

2.3.1.1 Embedded model based pre-selection

An embedded approach for feature pre-selection is achieved through the SVM classifier. The classification accuracy is evaluated for 246 properties using an open-source SVM implementation called LIBSVM [17] with a 10-fold cross validation on the training data set. Selection process is done on the basis of a threshold of 62%. All properties that obtained overall classification accuracy greater than or equal to the threshold are chosen. As a result, 145 properties are selected for further analysis.

2.3.1.2 Filter based pre-selection methods

The proposed filter model criterion is based on prior work by Zhou *et al.,* [24]. In their work, a modified t-test ranking measure is applied on HapMap genotype data. The same statistics is adopted in our study and evaluated for 246 properties. A feature i to be the greatest t-score for all classes, for feature i is given by

$$t_i = \max\left\{\frac{\left|\bar{x}_{ic} - \bar{x}_i\right|}{M_c S_i}\right\} \quad (2)$$

$$S_i^2 = \frac{1}{N-C}\sum_{c=1}^{C}\sum_{j\in c}\left(x_{ij} - \bar{x}_{ic}\right)^2 \quad (3)$$

$$M_c = \sqrt{1/n_c + 1/N} \quad (4)$$

Here t_i is the t-statistics value for the i^{th} property (feature); $\bar{x}_{ic}$ is the mean of the i^{th} feature in the c^{th} class, and $\bar{x}_i$ is the mean of the i^{th} feature for all classes; x_{ij} refers to the i^{th} feature of the j^{th} sample; N is the number of all the samples in the C classes and n_c is the number of samples in class c; S_i is the within-class standard deviation.

In addition, F-statistic test is utilized to perform feature pre-selection. Here, we computed the F-score [35] of each feature as follows: for a set of training vectors x_k, k=1, 2,---, m, if the number of positive and negative instances are n_+ and n_-, respectively, then the F-score of the i^{th} feature is defined as:

$$F(i) = \frac{\left(\bar{x}_i^{(+)} - \bar{x}_i\right)^2 + \left(\bar{x}_i^{(-)} - \bar{x}_i\right)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}\left(x_{k,i}^{(+)} - \bar{x}_i^{(+)}\right)^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}\left(x_{k,i}^{(-)} - \bar{x}_i^{(-)}\right)^2} \quad (5)$$

where $\bar{x}_i$ is the average value of feature i in all samples and $\bar{x}_i^{(+)}, \bar{x}_i^{(-)}$ are the average values of feature i in positive and negative samples respectively. To our knowledge, these statistical tests were not employed previously in literature to choose a subset of physicochemical properties of amino acids for the prediction of amyloid fibril segments. On applying the above statistics (modified t-test and F-score) independently, all 246 properties were ranked according to their scores, but it may be hard to decide how many top features should be selected into the final subset. However, to clearly assess the performances of filter and embedded methods, the same number of features selected with embedded SVM is chosen and hence the first top ranked 145 properties obtained by the statistical tests are considered for future analysis.

2.3.2 Wrapper of SVM for feature search

A wrapper method based on GA wrapped around the classifier, SVM to search for the significant minimal set of features is employed on 145 properties obtained from the pre-selection approaches separately.

The investigation on optimization techniques carried out by Kudo *et al*., [13] shows that the conventional search algorithms are the best for small and medium sized feature sets, while GA is better for large sized sets. This argument contradicts the findings in [14]. However, it is believed that one of the important factors affecting the GA results is due to the varying implementation of the GA method. In this study, one such variation of GA is adopted.

Initially, a set of randomly generated parents with 41 properties has been created. As mentioned [16], random selection is a selection operator where the best and worst individuals have exactly the same probability of surviving to the next generation. The parents were then ranked according to their fitness. To ensure that good individuals do survive to next generations, we chose the best half according to the fitness.

In order to produce offspring from the selected parents, a crossover operator and/or mutation operators are used [16]. Crossover could cause duplicates hence another way of recreation called 'Property pool' has been implemented for this task. All properties of the selected parents are put together in a pool. The offspring which replace the weaker parents have been built out of this pool. For every offspring, the properties were drawn one by one, saved if the property occurs the first time for the specific offspring else put back in to the pool. With this procedure it could be made sure, that a property which appeared more often in the fitter parents has a higher probability to be a part of the new generation.

Mutation makes sure that the properties, which are not part of the first generation, have a chance to get into the algorithm later. That means after the new offspring has been generated, there is a possibility that some properties of any offspring are replaced by randomly selected properties. A high mutation rate is desired for the first few generations, because it allows making big steps towards a better accuracy, but it should decrease with every generation to allow the algorithm to find the optimum with small changes. This method of mutation is called 'Simulated Annealing' and is commonly used in stochastic evolutionary algorithms [16]. To set the number of mutations per offspring, an exponential function has been implemented depending on the number of the actual generation.

$$N_M = s.m.e^{-6\frac{n_G}{N_G}} \quad (6)$$

where N_M is the number of mutations per offspring during the actual generation n_G, s is the size of offspring and m is the mutation value, a constant between 0 and 1 which defines the start value of the first generation depending on the size of an offspring. N_G stands for the total number of generations the algorithm is going to run for. In this work, the mutation value has been set to a value of 0.2 for every test.

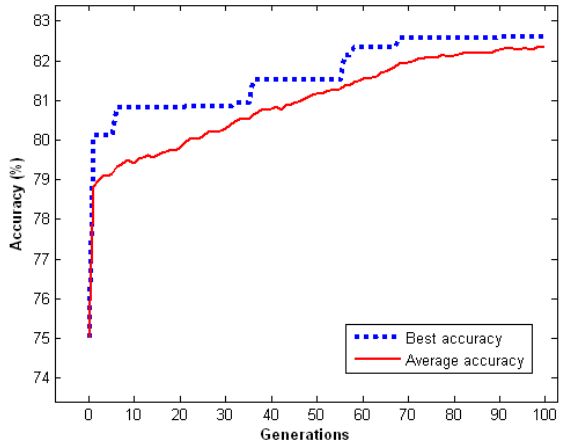

Figure 1. Average overall accuracies of all parents and the accuracies of best parents in different generations.

To calculate the fitness of a parent, LIBSVM [17] has been used which allows creating SVM and calculates the classification ability. LIBSVM comes with the *grid.py* script, an automatic grid search technique using cross-validation, which searches for the best parameters to use with the model training function of LIBSVM. Grid search was performed over the range $\log_2 C \in \{1, 2, 3\}$ and $\log_2 \gamma \in \{-3, -2, -1\}$. The SVM has been used with a Radial Basis Function (RBF-Kernel) and a 10-fold cross-validation.

The test was made with a generation size of 10 and 100 generations. As shown in figure 1, the classification ability/fitness of the parents becomes better with almost every generation. Recreation and mutation have not been creating better offspring all the time and because worse offspring are allowed in the GA, the overall accuracy of a generation could be lower than the accuracy of the previous generation. Nevertheless the accuracy has been increasing continuously because the best parents of every generation are kept in the algorithm and are just replaced if a new offspring has had a better classification ability. The 90th generation brought up the best combination of properties with an accuracy of 82.62%. The 41 properties obtained after feature selection are encoded for feature vector representation.

2.3.3 Integrated model for feature selection

In this work, we examine the performances of modified t-test [24] coupled with GA wrapper, F-statistic coupled with GA wrapper and embedded SVM classifier coupled with GA wrapper. The results of final prediction model indicate that the latter integration is better. Hence the chosen dimensionality reduction model is designed by employing an embedded SVM classifier for pre-selection integrated with GA wrapper method wrapped around SVM, with the goal of achieving leading-edge performance in predicting amyloidogenic peptides.

3. STATISTICAL ANALYSIS

Multiple measures were used to assess the performance of presented integrated FS methods including Sensitivity (S_n), Specificity (S_p), and Balanced Accuracy (BACC). In a binary classification, given a classifier and an instance, there are four possible outcomes [18]. When a positive instance is classified correctly as positive, it is counted as a true positive (TP); however if it is classified wrongly as negative, it is counted as a false negative (FN). If the instance is negative and has been classified correctly, it is counted as a true negative (TN), otherwise it is counted as a false positive (FP). In biomedical statistics, S_n is the probability of correctly predicting a positive example calculated as $S_n = \frac{TP}{TP + FN}$ and S_p is the chance of correctly predicting a negative example computed as $S_p = \frac{TN}{FP + TN}$. The BACC defines the arithmetic mean of S_n and S_p and is given by $BACC = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right)$. BACC is commonly used for evaluation of feature selection algorithms [34].

4. RESULTS AND DISCUSSION

It is widely recognized that a large number of features can adversely affect the performance of learning algorithms. This paper analyzed dimensionality reduction of bio-physio-chemical properties associated with amino acids using filter, wrapper, embedded approaches and integrated models to feature selection and compared their performances using SVM classifier based on statistical measures.

We tested the performances of integrated FS models using SVM classifier on a completely independent test dataset containing 1923 hexpeptides for which experimental data is available. We choose SVM classifier for it is a promising algorithm with high generalization ability, and is competitive with the best available learning machines in several applications including bioinformatics. Figure 2 depicts three measures illustrating the sensitivity and specificity, and the equilibrium maintained between them in terms of balanced prediction accuracy for the integrated FS models presented in this study. Filter methods (modified t-test and F-statistics) integrated with wrapper model, embedded method integrated with wrapper model and that without any feature selection result in a sensitivity of .80, .78 .86, .74 respectively and specificity of .77, .76, .82, .73 respectively. In addition to the maximum sensitivity and specificity scores obtained by embedded scheme integrated with wrapper approach, it has shown a good balance between sensitivity and specificity in terms of BACC (score of .84) in predicting a peptide status.

As evident, the prediction model trained with the features obtained by embedded SVM classifier coupled with GA wrapper tend to be superior to filter approaches coupled with GA wrapper and that without a feature selection. This could be due to the fact that filter methods attempt to select features based on simple auxillary criteria, such as feature correlation, to remove redundant features and simply rank individual features. In order to be tractable, such approaches decouple the feature selection process from the performance component, but may ultimately select irrelevant features as a result. Surprisingly, among the

univariate statistical tests of filter based methods considered in this study, modified t-test performs better than F-score test.

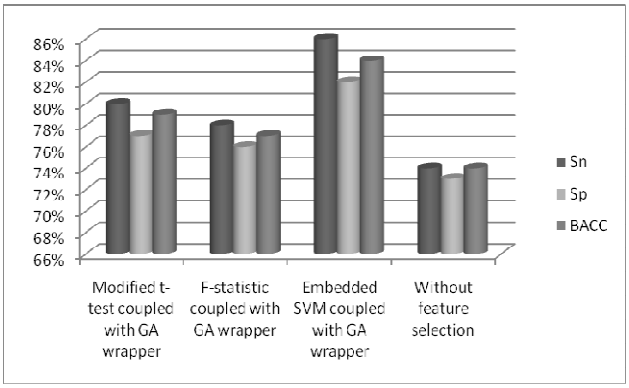

Figure 2. Comparative analysis of feature selection models in terms of sensitivity (S_n), specificity (S_p) and balanced prediction accuracy (BACC).

In order to truly assess the FS algorithms, it is imperative that datasets be free from flaws. In fact, experimentally predicted amyloidogenic regions reported in different works vary. One possibility could be due to the fact that the sequences are examined under diverse conditions. Hence reliable identification of amyloid fibril stretches is challenging and difficult. However, the effectiveness of preprocessing methods can be assessed only with respect to their ability to improve identification of relevant motif patterns governing class discrimination.

5. CONCLUDING REMARKS

The study of folding and unfolding events in proteins and subsequent aggregation into amyloid fibrillar deposits is becoming central to develop rational therapeutic strategies against maladies such as neurodegenerative diseases and Type II diabetes. A promising approach to spot such deposits is through computational prediction models. Due to the sheer amount of features contained within the amino acids, most standard machine learning algorithms cannot be directly applied. Instead, FS techniques are used to first reduce the dimensionality, thus enabling the subsequent use of classification methods effectively.

The aim of this paper is to profile a number of feature selection algorithms coupled with the SVM classifier. Our goal was to evaluate the performances in terms of statistical measures based on true positive rates, false positive rates and their balanced accuracy. To our knowledge, this is the first attempt to perform an extensive comparison between various FS categories on physicochemical properties of amino acids in predicting a peptide status: amyloidogenic or non-amyloidogenic. The present study examines the performance of wrapper method coupled with filter methods and that with embedded model. Moreover, a novel integrated approach has been designed resulting in a new and complementary set of physicochemical and biochemical properties to represent the feature vector. In addition, a variant of GA is implemented. To fairly assess each method, evaluation was done on the test dataset using SVM classifier that uses vector representations of sequences derived from selected sequence properties which revealed that embedded SVM classifier coupled with GA wrapper produce the most consistent results.

6. REFERENCES

[1] Amedeo Caflisch. 2007. Computational models for the prediction of polypeptide aggregation propensity, Current Opinion in Chemical Biology. ScienceDirect. 10: 437-444.

[2] Natalia Sánchez de Groot, Irantzu Pallarés, Francesc X Avilés, Josep Vendrell, and Salvador Ventura. 2005. Prediction of "hot spots" of aggregation in disease-linked polypeptides. BMC Structural Biology. 5:18, doi: 10. 1186/1472-6807-5-18.

[3] Amol P. Pawar, Kateri F. Dubay, Jesus Zurdo, Fabrizio Chiti, Michele Vendruscolo and Christopher M. Dobson. 2005. Prediction of "Aggregation-prone" and "Aggregation-susceptible" Regions in Proteins Associated with Neurodegenerative Diseases. J. Mol. Bio. 350, pp. 379-392.

[4] Jian Tian, Ningfeng Wu, Jun Guo and Yunliu Fan. 2009. Prediction of amyloid fibril-forming segments based on a support vector machine. BMC Bioinformatics. 10 (Suppl 1): S45.

[5] Manuela Lopez de la Paz and Luis Serrano. 2004. Sequence determinants of amyloid fibril formation. PNAS. Vol. 101, No. 1, pp. 87-92.

[6] Zhuqing Zhang, Hao Chen and Luhua La. 2007. Identification of amyloid fibril-forming segments based on structure and residue-based statistical potential. Structural Bioinformatics. Vol. 23 no. 17, pp. 2218–2225.

[7] Christopher J. C. Burges. 1998. A tutorial on Support Vector Machines for pattern recognition. Data Mining and Knowledge Discovery. 2(2), pp. 955-974.

[8] Michael J. Thompson, Stuart A. Sievers, John Karanicolas, Magdalena I. Ivanova, David Baker. 2006. The 3D profile method for identifying fibril-forming segments of proteins. PNAS. Vol. 103, No. 11, pp. 4074–4078.

[9] http://www.ebi.ac.uk/uniprot/database/download.html

[10] Kawashima S, Kanehisa M. 2008. AAindex: amino acid index database. Nucleic Acids Res. 28(1): 374.

[11] Jiawei Han, Micheline Kamber. 2008. Data Mining – Concepts and Techniques, Elsevier, II Edition.

[12] Laskko T A, Bhagwat J G, Zou K H, Ohno Machado L. 2005. The use of receiver operating characteristic curves in biomedical informatics. J Biomed Inform. 38: 404-415.

[13] Kudo M, Sklansky J. 2000. Comparison of algorithms that select features for pattern recognition. Pattern Recognition. 33(1): 25-41.

[14] Ferri F J, Pudil P, Hatef M, Kittler J. 1994. Comparative study of techniques for large-scale feature selection. Pattern Recognition in Practice IV, Elsevier. pp. 403-413.

[15] Yvan Saeys, Inaki Inza, Pedro Larran. 2007. A review of feature selection techniques in bioinformatics. Bioinformatics. Vol. 23 no. 19, pp. 2507–2517.

[16] Andries P. Engelbrecht. 2007. Computational Intelligence. John Wiley & Sons Ltd. Publishers, II Ed.

[17] http://www.csie.ntu.edu.tw/~cjlin/

[18] Pierre Baldi, Soren Brunak, Yves Chauvin, Claus A F Anderson, Henrick Nielson. 2000. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics. Vol 16, No. 5.

[19] Smitha Sunil Kumaran Nair, N. V. Subba Reddy, Hareesha K. S. 2010. Computational models for the prediction of amyloid fibril forming protein segments. Proc. Int'l Conference on Bioinformatics and Systems Biology, Annamalai University, Vol. 1, pp.152-157.

[20] Kimon K Frousios, Vassiliki A Iconomidou, Carolina-Maria Karletidi, Stavros J Hamodrakas. 2009. Amyloidogenic deteminants are usually not buried. BMC Structural Biology. 9:44.

[21] Oxana V. Galzitskaya, sergiy O. Garbuzynskiy, Michail Yurievich Lobanov. 2006. Prediction of Amyloidigenic and Disordered Regions in Protein Chains. PLoS Computational Biology. Volume 2, Issue 12, e177.

[22] Magdalena I. Ivanova, Michael J. Thompson, and David Eisenberg. 2006. A systematic screen of β2-microglobulin and insulin for amyloid-like segments. PNAS. Vol. 103, No. 11, pp. 4079–4082.

[23] Ana-Maria Fernandez-Escamilla, Frederic Rousseau, Joost Schymkowitz & Luis Serrano. 2004. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. Nature Biotechnology. Vol. 22, No. 10, pp. 1302-1306.

[24] Nina Zhou and Lipo Wang. 2007. A Modified T-test Feature Selection Method and Its Application on the HapMap Genotype Data: Geno. Prot. Bioinfo. Vol. 5. No. 3-4, pp. 242-249.

[25] http://antares.protres.ru/fold-amyloid/

[26] Oscar Conchillo-Sole, Natalia S de Groot, Francesc X Aviles, Josep Vendrell, Xavier Daura and Salvador Ventura. 2010. AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. BMC Bioinformatics. 8:65.

[27] Susan Idicula-Thomas and Petety V Balaji. 2005. Understanding the relationship between the primary structure of proteins and their amyloidogenic propensity: clues from inclusion body formation. Journal of Protein Engineering, Design & Selection. Vol. 18, No. 4, pp. 175-180.

[28] Mathura & Kolippakkam. 2005. APDbase: Amino acid Physicochemical properties Database. Bioinformation. 1(1): 2-4.

[29] http://www.expasy.org/tools/protscale.html

[30] http://www.rfdn.org/bioinfo/APDbase.php

[31] Sergiy O. Garbuzynskiy, Michail Yu. Lobanov and Oxana V. Galzitskaya. 2010. FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. Structural Bioinformatics. Vol. 26, No.3, pp.326-332.

[32] http://biophysics.biol.uoa.gr/AMYLPRED/input.html

[33] Sukjoon Yoon, William J. Welsh. 2004. Detecting hidden sequence propensity for amyloid fibril formation. Protein Science. 13: 2149-2160.

[34] Ilya Levner. 2005. Feature selection and nearest centroid classification for protein mass spectrometry. BMC Bioinformatics. 6:68.

[35] Sanghamitra Bandyopadhyay, Ramkrishna Mitra. 2009. TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative example. Bioinformatics. Vol. 25, No. 20, pp. 2625-2631.

[36] Shipin Lv, Xiukun Wang, Yifen Cui, Jue Jin, Yan Sun, Yiyuan Tang, Ying Bai, Yan Wang, Li Zhou. 2010. Application of attention network test and demographic information to detect mild cognitive impairment via combining feature selection with support vector machine. Computer Methods and programs in Biomedicine 97, Elsevier. pp. 11-18.