Published in final edited form as:
*Mol Biosyst*. 2013 April 5; 9(4): 806–811. doi:10.1039/c3mb70033j.

# Discrimination of soluble and aggregation-prone proteins based on sequence information

**Yaping Fang** and **Jianwen Fang***
Applied Bioinformatics Laboratory, The University of Kansas, 2034 Becker Dr., Lawrence, Kansas 66047, USA

## Abstract

Understanding the factors governing protein solubility is a key to grasp the mechanisms of protein solubility and may provide insight into protein aggregation and misfolding related diseases such as Alzheimer's disease. In this work, we attempt to identify factors important to protein solubility using feature selection. Firstly, we calculate 1438 features including physicochemical properties and statistics for each protein. Random Forest algorithm is used to select the most informative and the minimal subset of features based on their predictive performance. A predictive model is built based on 17 selected features. Compared with previous models, our model achieves better performance with a sensitivity of 0.82, specificity 0.85, ACC 0.84, AUC 0.91 and MCC 0.67. Furthermore, a model using redundancy-reduced dataset (sequence identity <= 30%) achieves the same performance as the model without redundancy reduction. Our results provide not only a reliable model for predicting protein solubility but also a list of features important to protein solubility. The predictive model is implemented as a freely available web application at http://shark.abl.ku.edu/ProS/.

### Keywords

Protein solubility; Aggregation; Random Forest; Classification; Feature selection

## Introduction

Protein solubility plays an important role in protein production and application [1–3]. It was estimated that ca. 33–50% of all expressed non-membrane proteins are insoluble, and ca. 25–57% of those soluble proteins prone to aggregate or precipitate at higher concentrations [4–6]. Thus understanding the factors governing protein solubility is important to grasp the mechanisms of protein solubility and improve the efficiency of designing soluble proteins. Moreover, it may provide insight into protein aggregation and misfolding related diseases such as Alzheimer's disease [7, 8].

Exiting predictive methods on protein solubility can be generally grouped into two distinct classes: structure-based and sequence-based. The structure-based methods usually calculate the free energy difference between solution and aggregation phases [9, 10]. This type of methods requires experimentally-determined high resolution three-dimensional structures, which can be difficult to obtain for aggregation prone proteins. Thus often only sequence-based approaches are feasible. A number of sequence-based methods have been already developed. For example, Wilkinson and Harrison analyzed 81 proteins and found that protein solubility is related to amino acid composition [11]. A revised Wilkinson–Harrison

*To whom correspondence should be addressed. jwfang@ku.edu.

solubility model was later published by Davis *et al.* [12]. Idicula-Thomas and Balaji also found that amino acid composition especially the proportion of Asn, Thr and Tyr residues and other sequence-dependent features have impact on solubility of over-expressed proteins [13]. They developed a Support Vector Machine (SVM) predictive model based on six physicochemical properties and amino acid composition of 192 protein sequences including 130 insoluble and 62 soluble proteins [14]. The six physicochemical properties include length of protein, hydropathic index, aliphatic Index, instability index, instability index of N-terminus and net charge. Another study, however, revealed that sequence-based methods based on small dataset may have poor generalization ability [15].

Recently, several methods have been built on larger datasets. For example, Smialowski *et al.* built a model named PROSO based on 14000 protein sequences [16]. The model achieved an accuracy of 72% in their tests [16]. More recently, the same group of authors reported an improved model, PROSOII, achieving an accuracy of 75.4% using a logistic function and an adapted Parzen window algorithm based on k-mer properties of 82000 proteins [17]. Magnan *et al.* constructed a SVM model SOLpro with 74% accuracy based on 17000 protein sequences and the frequencies of monomers, di-mers and tri-mers of amino acids [18]. It should be pointed out that the datasets used to build PROSO, PROSOII and SOLpro were collected by incorporating different search results of Protein Data Bank (PDB) [15], Swiss-Prot database and TargetDB [19]. The proteins were then classified into soluble and insoluble ones based on the annotations of these proteins. While these approaches were best practices when a suitable experimental dataset was not available, they may not be always reliable. For example, a soluble protein missing proper annotation can be mistakenly classified as an insoluble one, and *vice versa*. In addition, annotations from different databases are not consistent with each other. Obviously a large set of protein with experimentally determined solubility using a single consistent protocol is desirable.

Recently, Niwa *et al.* analyzed the solubility of entire proteome of Escherichia coli (*E. Coli*) using a cell-free translation system and classified the proteins into soluble and aggregation-prone proteins[20]. The authors built a predictive model using the SVM algorithm based on molecular weight, isometric point (pI) and amino acid composition which achieved ~80% accuracy. Using this dataset, Stiglic *et al.* developed a comprehensive decision tree model to classify the soluble and aggregation-prone proteins based on the sequence information[21]. This model achieves an accuracy of 72 % based on a 10-fold cross validation. Both studies have revealed that amino acid composition, molecular weight and pI of proteins are relevant to protein solubility. However, there is little systematic investigation on the relative importance of various types of features used to build reliable models. Thus the goal of this study is to build a model for predicting protein solubility using the most informative and minimal subset features identified using a state-of-the-art feature selection algorithm. Such a study can provide information for not only accurately predicting protein solubility but also aiding in discovering underlying mechanisms of protein solubility.

## Materials and methodsd

### Datasets

All proteins used in the study were downloaded from eSOL database (http://tp-esol.genes.nig.ac.jp/)[20] in February 2012. Only proteins with available sequences are retained. A protein with solubility < 30% is considered as aggregation-prone and a protein with solubility >70% is considered as soluble[20]. There are 2183 proteins, including 988 soluble and 1195 aggregation-prone proteins. We then prepare a series of subsets with the sequence identities no higher than 90%, 75%, 50% and 30% using the CD-Hit program[22]. We use the set of 30% identity, including 1918 proteins (886 soluble and 1032 aggregation-prone proteins), to build the final model.

### Features

Each protein is encoded with 1438 features that can be grouped into four classes (Table 1). The first class (I) is physicochemical properties which are the average values of amino acids for a given protein. The second class (II) includes absolute counts and normalized absolute counts by the length of amino acids for a given protein. The third class (III) is absolute counts and normalized absolute counts by the protein length of di-peptide for a given protein. The fourth class (IV) includes the remaining features. All 1438 features are sequence-based features or structural features which are predicted from sequences. Although actual structural information should be useful in predicting protein solubility, most proteins in eSOL database have no solved structures. In addition, previous studies[20, 21] have revealed that sequence-dependent features can be effective in predicting protein solubility.

### Random Forest

The Random Forest (RF) algorithm [32] is an ensemble machine learning method that utilizes many independent decision trees to perform classification or regression. Each of the member trees is built on bootstrap samples from the training data by a random subset of available variables. RF models built in this study consist of 5000 decision trees. The number of variable randomly sampled in each tree is $\sqrt{M}$, where M is the number of total variables. The RF algorithm has been successfully used in a number of predictive models [33–35]. An important application of the algorithm is to assess the importance of various features based on their contributions to the performance. In this study we used variable importance of features which is based on the mean decrease in accuracy [32]. An R package varSelRF utilizing feature importance for feature selection is used to identify the most informative and minimal subset features [36].

### Performance assessment

Several metrics are used to quantitatively assess the performance. The receiver operation characteristic (ROC) curve is a graphic plot of the true-positive rate (sensitivity) against the false-positive rate (1-specificity). The area under an ROC curve (AUC) shows the trade-off between sensitivity and specificity. The value of AUC is in the range of 0 to 1 and the bigger the AUC, the better the performance is. An AUC of 0.5 represents random classification and 1 is for perfect prediction. The other metrics include:

$$\text{Sensitivity}=\frac{\text{TP}}{\text{TP+FN}} \quad (2)$$

$$\text{Specificity}=\frac{\text{TN}}{\text{TN+FP}} \quad (3)$$

$$\text{MCC}=\frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (4)$$

$$\text{ACC}=\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (5)$$

Where TP, TN, FP and FN are true positive, true negative, false positive and false negative respectively. The MCC is a measure of a correlation coefficient between the observed and predicted binary classifications. It has value between −1 and +1. A coefficient of +1 represents a perfect prediction, 0 no better than random prediction and −1 indicates total

disagreement between prediction and observation. ACC with 100% represent perfect classification model. And the bigger the ACC, the better the performance.

## Results and Discussion

### Amino Acid composition

The amino acid composition of both soluble and aggregation-prone proteins is shown in Table 2. The statistical difference of amino acid composition between soluble and aggregation-prone proteins is estimated using the student *t*-test. The rows in Table 2 highlighted in red or blue are amino acid residues significantly over-represented (*p*-value < $10^{-5}$) in soluble and aggregation-prone proteins, respectively. It can be seen that aggregation-prone proteins tend to have more Serine (S), Tyrosine (Y), Phenylalanine (F), Leucine (L), Proline (P), Tryptophan (W) and Arginine (R) residues, indicating that a protein with more aromatic amino acid residues tend to be an aggregation-prone protein. This finding is consistent with previous findings [5, 17]. The soluble proteins tend to have more Aspartic (D), Glutamic (E) and Lysine (K) residues, indicating that soluble proteins tend to have more charged residues than aggregation-prone proteins. Interestingly, Leucine, Isoleucine and valine, three amino acids with considerably similar physicochemical properties, are distributed significantly different in soluble and aggregation-prone proteins. While the content of isoleucine residue reminds largely unchanged in these two groups, leucine residue is significantly enriched in aggregation-prone proteins and leucine residue instead enriched in the other group. Overall, charged and aromatic amino acid residues are important to protein solubility.

### Performance of the feature sets

To estimate the relative importance and relevance of feature sets to the solubility, we build a series of models using different combinations of the four feature sets (Table 3). The model using all features achieves the best performance, suggesting all features are relevant to protein solubility to some extent. Different feature groups have different ability in classifying soluble and aggregation-prone proteins. The amino acid composition features are most important and the dipeptide features are least important. Although overall the dipeptide features may contain more information than the amino acid composition, the information density for each dipeptide is probably very low because there are 400 dipeptides. Thus the importance of each individual dipeptide is low.

### Features importance

To select the most informative and minimal subset features, the varSelRF package [36] is used to iteratively eliminate 10% features for each iteration. Two approaches are used to select features. One way re-calculates the importance of features after each iteration and the second only evaluates the importance once. The first method results in 17 features (F17). For the second method, the features are firstly sorted as descending order of importance and the top17 features are selected (FI17). F17 and FI17 share 16 features.

The annotations and the relative importance of F17 features are listed in Table 4 and Figure 1 respectively. The top 5 features with highest variable importance include free energies of transfer (WIMW960101), partition coefficient (ZASB820101) of proteins, net charge of protein (x_netcharge and KLEP840101) and isometric point (pI), consistent with previous studies that pI [20, 21] and free energies of transfer [9, 10] are important to classify soluble and aggregation-prone proteins. Partition coefficient is an important parameter related to molecular solubility. Our study also identifies other features important to protein solubility such as aromatic amino acid content, surface composition of amino acids, nitrogen atom content, beta-strand indices for beta-proteins and hydrogen bond. Compared to protein

length which is important to protein solubility [37], the number of nitrogen atom reveals more specific information. The results also indicate that the contents of amino acid R and L are important to protein solubility. Amino acid R is important to maintain the overall charge balance of proteins and amino acid L is generally buried in proteins. It is also consistent with amino acid composition analysis presented previously. Partition coefficient, solvent accessible surface area and the number of buried amino acids are also important factors that have influence on protein solubility.

### Performance with the sequence identity

To further evaluate the effectiveness of selected features, several models have been rebuilt at different sequence identity. Specifically, the sequence identity of proteins is reduced to different levels such as 90%, 75%, 50% and 30% using the CD-Hit program [22]. The selected features are then used to rebuild the models. The number of protein sequences and the performance at different sequence identity levels are shown in Table 5. Both F17- and FI17-based models achieve good performance at various sequence identity levels. The performance of models based on F17 features is slightly better than or equivalent to those of models based on FI17 features. Thus, the features F17 are used in the final model. It can be seen that although the sequence identity is reduced to 30%, the model still has performance with sensitivity 0.82, specificity 0.85, ACC 0.84, AUC 0.91 and MCC 0.67. The results indicate that the selected features are effective and can be applied to build the models based on both strict and loose sequence identity.

### Comparing to previous methods

There are several previous methods to predict protein solubility based on eSOL database such as SVM model [20], decision tree model (VTJ48 and J48) [21]. Their performances are shown in Table 5. Table 5 shows that the SVM model [20] achieved an accuracy of ~0.80, the VTJ48 model resulted in an accuracy of 0.76 and J48 had an accuracy of 0.72. The results indicate that our method has the best performance with sensitivity 0.82, specificity 0.85, ACC 0.84, AUC 0.91 and MCC 0.67. The SVM model [20] is a close second.

## Conclusions

Protein solubility plays an important role in various fields such as pharmacy, food and protein storage. In this work, we use the RF algorithm on a unified experimental verified protein dataset in eSOL database (http://tp-esol.genes.nig.ac.jp/) to identified 17 features which are important to protein solubility. Besides some features are consistent with previous works such as the number of aromatic amino acids, negative charge amino acids, PI and transfer free energy [2, 17, 18, 20, 21], several new features are found such as partition coefficient, solvent accessible surface area, the number of buried amino acids, long range non-bonded energy, beta-strand indices and flexibility parameter. Based on such 17 features, a predictive model using RF algorithm is built. Compared with existing methods on the same dataset, it has the best performance with sensitivity 0.82, specificity 0.85, ACC 0.84, AUC 0.91 and MCC 0.67. The results indicate that selected features can be effectively used in discriminating soluble proteins and aggregation-prone proteins. The built model and subset of selected sequence features should have roles in soluble protein design. The final predictive model is implemented as a freely available web application at http://shark.abl.ku.edu/ProS/.

## Acknowledgments

We wish to thank the two anonymous reviewers for their constructive comments and suggestions. This work was supported in part by the National Institutes of Health (NIH) Grant P01 AG12993 (PI: E. Michaelis).

## References


1. Pace CN, Trevino S, Prabhakaran E, Scholtz JM. Philos Trans R Soc Lond B Biol Sci. 2004; 359:1225–1234. discussion 1234-1225. [PubMed: 15306378]
2. Tjong H, Zhou HX. Biophys J. 2008; 95:2601–2609. [PubMed: 18515380]
3. Mandava N, Oberoi RK, Minocha M, Mitra AK. J Drug Deliv Sci Tec. 2010; 20:89–99.
4. Yee A, Pardee K, Christendat D, Savchenko A, Edwards AM, Arrowsmith CH. Accounts of chemical research. 2003; 36:183–189. [PubMed: 12641475]
5. Christendat D, Yee A, Dharamsi A, Kluger Y, Savchenko A, Cort JR, Booth V, Mackereth CD, Saridakis V, Ekiel I, Kozlov G, Maxwell KL, Wu N, McIntosh LP, Gehring K, Kennedy MA, Davidson AR, Pai EF, Gerstein M, Edwards AM, Arrowsmith CH. Nature structural biology. 2000; 7:903–909.
6. Yee A, Chang X, Pineda-Lucena A, Wu B, Semesi A, Le B, Ramelot T, Lee GM, Bhattacharyya S, Gutierrez P, Denisov A, Lee CH, Cort JR, Kozlov G, Liao J, Finak G, Chen L, Wishart D, Lee W, McIntosh LP, Gehring K, Kennedy MA, Edwards AM, Arrowsmith CH. Proc Natl Acad Sci U S A. 2002; 99:1825–1830. [PubMed: 11854485]
7. Woltjer RL, Montine TJ. Faseb Journal. 2006; 20:A1088–A1088.
8. Vendruscolo M, Knowles TPJ, Dobson CM. Csh Perspect Biol. 2011; 3
9. Tjong H, Zhou HX. Biophys J. 2008; 95:2601–2609. [PubMed: 18515380]
10. Ahmad SS, Dalby PA. Biotechnol Bioeng. 2011; 108:322–332. [PubMed: 20872822]
11. Wilkinson DL, Harrison RG. Biotechnology (N Y). 1991; 9:443–448. [PubMed: 1367308]
12. Davis GD, Elisee C, Newham DM, Harrison RG. Biotechnology and Bioengineering. 1999; 65:382–388. [PubMed: 10506413]
13. Idicula-Thomas S, Balaji PV. Protein Science. 2005; 14:582–592. [PubMed: 15689506]
14. Idicula-Thomas S, Kulkarni AJ, Jayaraman VK, Balaji PV. Bioinformatics (Oxford, England). 2006; 22:278–284.
15. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C. Acta Crystallogr D Biol Crystallogr. 2002; 58:899–907. [PubMed: 12037327]
16. Smialowski P, Martin-Galiano AJ, Mikolajka A, Girschick T, Holak TA, Frishman D. Bioinformatics (Oxford, England). 2007; 23:2536–2542.
17. Smialowski P, Doose G, Torkler P, Kaufmann S, Frishman D. FEBS J. 2012
18. Magnan CN, Randall A, Baldi P. Bioinformatics (Oxford, England). 2009; 25:2200–2207.
19. Chen L, Oughtred R, Berman HM, Westbrook J. Bioinformatics (Oxford, England). 2004; 20:2860–2862.
20. Niwa T, Ying BW, Saito K, Jin W, Takada S, Ueda T, Taguchi H. Proceedings of the National Academy of Sciences of the United States of America. 2009; 106:4201–4206. [PubMed: 19251648]
21. Stiglic G, Kocbek S, Pernek I, Kokol P. PLoS One. 2012; 7:e33812. [PubMed: 22479449]
22. Huang Y, Niu B, Gao Y, Fu L, Li W. Bioinformatics (Oxford, England). 2010; 26:680–682.
23. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. Nucleic acids research. 2008; 36:D202–D205. [PubMed: 17998252]
24. Galzitskaya OV, Garbuzynskiy SO, Lobanov MY. PLoS Comput Biol. 2006; 2:e177. [PubMed: 17196033]
25. Conchillo-Sole O, de Groot NS, Aviles FX, Vendrell J, Daura X, Ventura S. Bmc Bioinformatics. 2007; 8
26. Pawar AP, DuBay KF, Zurdo J, Chiti F, Vendruscolo M, Dobson CM. Journal of molecular biology. 2005; 350:379–392. [PubMed: 15925383]
27. Chennamsetty N, Voynov V, Kayser V, Helk B, Trout BL. J Phys Chem B. 2010; 114:6614–6624. [PubMed: 20411962]

28. Tartaglia GG, Cavalli A, Pellarin R, Caflisch A. Protein Sci. 2005; 14:2723–2734. [PubMed: 16195556]
29. Eisenhaber F, Argos P. Journal of Computational Chemistry. 1993; 14:1272–1280.
30. Wilkins MR, Gasteiger E, Bairoch A, Sanchez JC, Williams KL, Appel RD, Hochstrasser DF. Methods Mol Biol. 1999; 112:531–552. [PubMed: 10027275]
31. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. Structure. 2003; 11:1453–1459. [PubMed: 14604535]
32. Breiman L. Machine Learning. 2001; 45:5–32.
33. Sikic M, Tomic S, Vlahovicek K. PLoS Comput Biol. 2009; 5:e1000278. [PubMed: 19180183]
34. Wang L, Yang MQ, Yang JY. BMC Genomics. 2009; 10(Suppl 1):S1. [PubMed: 19594868]
35. Li Y, Fang Y, Fang J. Bioinformatics (Oxford, England). 2011; 27:3379–3384.
36. Diaz-Uriarte R. Bmc Bioinformatics. 2007; 8:328. [PubMed: 17767709]
37. Goh CS, Lan N, Douglas SM, Wu B, Echols N, Smith A, Milburn D, Montelione GT, Zhao H, Gerstein M. Journal of molecular biology. 2004; 336:115–130. [PubMed: 14741208]
38. Fauchere JL, Charton M, Kier LB, Verloop A, Pliska V. Int J Pept Protein Res. 1988; 32:269–278. [PubMed: 3209351]
39. Fukuchi S, Nishikawa K. Journal of molecular biology. 2001; 309:835–843. [PubMed: 11399062]
40. Geisow MJ, Roberts RDB. International Journal of Biological Macromolecules. 1980; 2:387–389.
41. Karplus PA, Schulz GE. Naturwissenschaften. 1985; 72:212–213.
42. Klein P, Kanehisa M, DeLisi C. Biochim Biophys Acta. 1984; 787:221–226. [PubMed: 6547351]
43. Oobatake M, Ooi T. J Theor Biol. 1977; 67:567–584. [PubMed: 904331]
44. Wimley WC, White SH. Nature Structural Biology. 1996; 3:842–848.
45. Zaslavsky BY, Mestechkina NM, Miheeva LM, Rogozhin SV. J Chromatogr. 1982; 240:21–28.

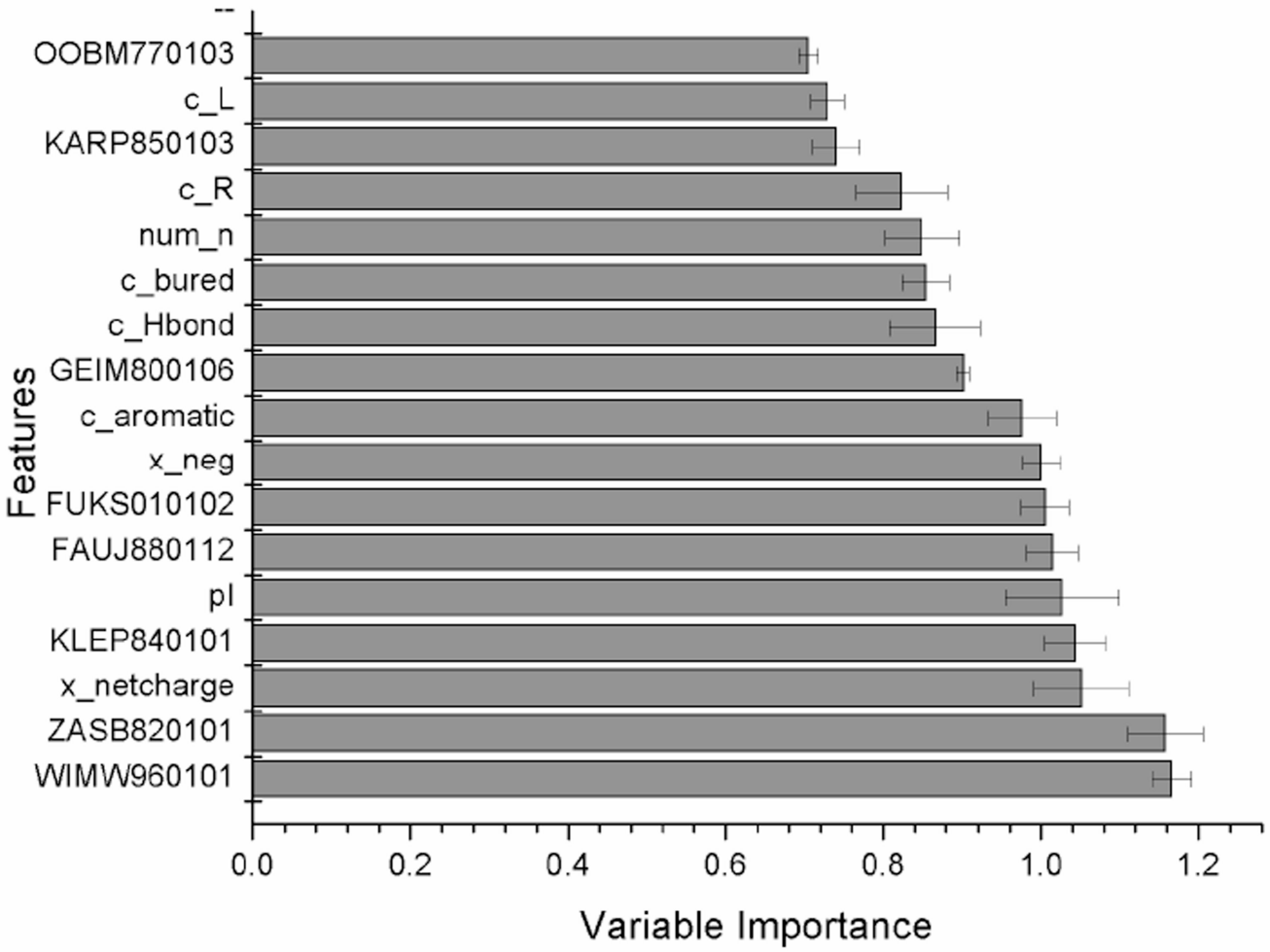


**Figure 1. Variable importance of F17 features**

The prefix x represents the normalized absolute count values and c represents the absolute count values for each amino acid. The prefix num means the count of a specific atom. The other features are physicochemical properties of AAindex database.

## Table 1

The list of 1438 sequence dependent features

| Group | Protein features | Number of Features | Source |
|---|---|---|---|
| I | physicochemical properties obtained from AAindex | 544 | [23] |
| | Density | 1 | [24] |
| | Relative experimental aggregation propensities | 1 | [25] |
| | Amyloid aggregation propensities | 1 | [26] |
| | Solvent accessible area of exposed side chains | 1 | [27] |
| | Property index | 12 | [28] |
| II | Number and composition of amino acids | 40 | In house script |
| III | Number and composition of dipeptides | 800 | |
| IV | Sequence length (L) | 1 | |
| | Number and percentage of positive, negative and all charged residues, as well as the net charges | 8 | |
| | Number and percentage of small (T and D), tiny (G, A, S and P), aromatic (F, H, Y, W), aliphatic, hydrophobic and polar residues | 12 | |
| | Number and percentage of residues which can form hydrogen bond in side chain | 2 | |
| | The average of the maximum solvent accessible surface area (ASA) of each amino acid | 1 | Eisenhaber [29] |
| | Predicted isoelectric point (pI) of protein, the average pI on all residues (pIa) | 2 | ProtParam [30] |
| | Instability index and instability class | 2 | |
| | Aliphatic index | 1 | |
| | Gravy hydropathy index | 1 | |
| | The overall length and percentage of all coils, rem465, and hotloop | 6 | disEMBL [31] |
| | Mean Relative Surface Accessibility - RSA | 1 | |
| | Mean Z-fit score for RSA prediction | 1 | |

## Table 2

Amino acid composition of soluble and aggregation-prone proteins

| Amino Acid | Composition in soluble proteins | Composition in aggregation-prone proteins | *P*-value (*t*-test) |
|---|---|---|---|
| **S** | **0.054±0.021** | **0.059±0.017** | **2.52e-09** |
| Q | 0.046±0.022 | 0.046±0.018 | 0.43 |
| N | 0.040±0.018 | 0.039±0.017 | 0.60 |
| T | 0.054±0.022 | 0.052±0.015 | 0.097 |
| C | 0.013±0.014 | 0.013±0.011 | 0.86 |
| G | 0.070±0.027 | 0.071±0.021 | 0.10 |
| A | 0.095±0.031 | 0.09237±0.025 | 0.026 |
| H | 0.023±0.015 | 0.025±0.012 | 6.11e-05 |
| M | 0.028±0.013 | 0.029±0.011 | 0.44 |
| **Y** | **0.025±0.014** | **0.030±0.014** | **1.31e-15** |
| **F** | **0.032±0.016** | **0.039±0.015** | **1.41e-23** |
| **V** | **0.072±0.023** | **0.067±0.019** | **7.92e-06** |
| **L** | **0.097±0.029** | **0.11±0.029** | **6.41e-25** |
| **P** | **0.042±0.019** | **0.045±0.015** | **6.33e-06** |
| I | 0.058±0.020 | 0.059±0.020 | 0.19 |
| **W** | **0.011±0.010** | **0.017±0.011** | **4.91e-32** |
| **D** | **0.059±0.019** | **0.049±0.016** | **1.02e-35** |
| **E** | **0.072±0.026** | **0.055±0.019** | **1.23e-56** |
| **K** | **0.056±0.028** | **0.041±0.018** | **1.96e-46** |
| **R** | **0.055±0.026** | **0.06±0.019** | **3.23e-06** |

The *p*-values are based on *t*-test. Amino acids significantly increased or reduced in soluble proteins than aggregation-prone proteins are colored in red or blue, respectively.

**Table 3**

Performance of models built on different feature groups

| Group | Sensitivity | Specificity | ACC | MCC |
|---|---|---|---|---|
| I | 0.79 | 0.82 | 0.81 | 0.61 |
| II | 0.80 | 0.85 | 0.82 | 0.65 |
| III | 0.67 | 0.82 | 0.75 | 0.50 |
| IV | 0.78 | 0.85 | 0.82 | 0.64 |
| I+II | 0.83 | 0.84 | 0.83 | 0.66 |
| I+III | 0.83 | 0.84 | 0.84 | 0.67 |
| I+IV | 0.83 | 0.85 | 0.84 | 0.68 |
| II+III | 0.70 | 0.85 | 0.79 | 0.55 |
| II+IV | 0.81 | 0.86 | 0.83 | 0.67 |
| III+IV | 0.73 | 0.86 | 0.80 | 0.60 |
| I+II+III | 0.83 | 0.85 | 0.84 | 0.68 |
| I+II+IV | 0.82 | 0.85 | 0.84 | 0.68 |
| I+III+IV | 0.83 | 0.85 | 0.84 | 0.68 |
| II+III+IV | 0.77 | 0.86 | 0.82 | 0.63 |
| I+II+III+IV | 0.83 | 0.85 | 0.84 | 0.68 |

The feature sets are physicochemical properties (I), amino acid features (II), di-peptide features (III) and other features (IV).

## Table 4

Selected features and annotations of F17

| Feature | Annotation |
|---|---|
| c_aromatic | Counts of aromatic amino acids |
| c_bured | Counts of buried amino acids |
| c_Hbond | Counts of hydrogen bonds |
| c_L | Counts of leucine amino acid |
| c_R | Counts of arginine amino acid |
| FAUJ880112 | Negative charge [38] |
| FUKS010102 | Surface composition of amino acids in intracellular proteins of mesophiles (percent) [39] |
| GEIM800106 | Beta-strand indices for beta- proteins [40] |
| KARP850103 | Flexibility parameter for two rigid neighbors [41] |
| KLEP840101 | Net charge [42] |
| num_n | Counts of nitrogen atoms |
| OOBM770103 | Long range non-bonded energy per atom [43] |
| pI | Isometric point |
| WIMW960101 | Free energies of transfer of AcWl-X-LL peptides from bilayer interface to water [44] |
| x_neg | Ratio of negative charge amino acids |
| x_netcharge | Ratio of net charge of protein |
| ZASB820101 | Dependence of partition coefficient on ionic strength [45] |

The prefix x represents the normalized absolute count values and c represents the absolute count values for each amino acid. The prefix num means the count of a specific atom. The other features are physicochemical properties of AAindex database.

**Table 5**

The numbers of protein sequences and the performance of different models for different sequence identity

| sequence identity | # of solubility proteins | # of aggregation-prone proteins | Sensitivity | Specificity | ACC | AUC | MCC |
|---|---|---|---|---|---|---|---|
| FI17 features | | | | | | | |
| 90% | 983 | 1192 | 0.81 | 0.85 | 0.83 | 0.91 | 0.66 |
| 75% | 978 | 1189 | 0.81 | 0.85 | 0.83 | 0.91 | 0.66 |
| 50% | 968 | 1158 | 0.81 | 0.85 | 0.83 | 0.91 | 0.67 |
| 30% | 886 | 1032 | 0.82 | 0.85 | 0.84 | 0.91 | 0.67 |
| F17 features | | | | | | | |
| 90% | 983 | 1192 | 0.82 | 0.85 | 0.84 | 0.91 | 0.67 |
| 75% | 978 | 1189 | 0.81 | 0.85 | 0.84 | 0.91 | 0.67 |
| 50% | 968 | 1158 | 0.82 | 0.85 | 0.84 | 0.91 | 0.67 |
| 30% | 886 | 1032 | 0.82 | 0.85 | 0.84 | 0.91 | 0.67 |

**Table 5**

Comparison of our model to previous reported methods

| Method | Sensitivity | Specificity | ACC | AUC | MCC |
|---|---|---|---|---|---|
| SVM [20] | - | - | ~0.80 | - | - |
| VTJ48 [21] | - | - | 0.76 | 0.81 | - |
| J48[21] | - | - | 0.72 | 0.72 | - |
| Our method | **0.82** | **0.85** | **0.84** | **0.91** | **0.67** |