Esmaeil Ebrahimie, Mansour Ebrahimi, Mahdi Ebrahimi
Amino acid features: a missing compartment of prediction of protein function
Nature Precedings, 2011; Online:1-16

Originally published at:
http://doi.org/10.1038/npre.2011.6693.1

Nature Precedings

Title

Amino acid features: a missing compartment of prediction of protein function

Esmaeil Ebrahimie[1*], Mansour Ebrahimi[2], Mahdi Ebrahimi[3]

[1*]Molecular and Biomedical Science, The University of Adelaide, Adelaide, Australia

[2]Department of Biology & Bioinformatics Research Group, University of Qom, Iran

[3]Max-Planck-Institute for Informatics, Saarbrucken, Germany

Email: esmaeil.ebrahimie@adelaide.edu.au

Abstract

Enormous computational efforts have been carried out to predict structure and function of protein. However, nearly all of these efforts have been focused on prediction of function based on primary nucleic acid sequence or modelling 3D structure of protein from its nucleic acid sequence. In fact, it seems that amino acid attributes, which is an intermediate phase between DNA/RNA and advanced protein structure, have been missed.

From 2010, we examined the possibility of precise prediction of structural protein function based on amino acid features by improving the following three aspects of amino acid research: (1) Increasing the number of computationally calculated amino acid features, (2) Testing different feature selection (attribute weighting) algorithms and selection of the most important amino acid attribute based on the overall conclusion of algorithms, (3) Examining different supervised and unsupervised data mining (machine learning) algorithms, and (4) Joining attribute weighting with different data mining algorithms. We applied the discovered procedure in different biological examples including: protein thermostability, halostability, prediction of function of heavy metal transporters, cancer diagnosis and prediction, and pursuing the EST-SSRs in amino acid level.

In thermostability study, we successfully established an accurate expert system to predict the thermostability of any input sequence trough mining of its calculated amino acid features. Interestingly, performance of a clustering algorithm such as

EMC can vary from 0.0% to 100%, depending upon which attribute weighting algorithm had summarized the attributes of the dataset prior to running the clustering algorithm.

In another recent study on halostability, the results showed that amino acid composition can be used to efficiently discriminate halostable protein groups with up to 98% accuracy implying the possibility of precise prediction of halostability when an appropriate machine learning algorithm mines a large number of structural amino acid attributes of primary protein structure.

Using our approach, simple amino acid features, without the need of advanced features of protein structure, could explain the difference between P1B-ATPases in hyperaccumulator and nonhyperaccumulator plants. More importantly, a precise model was built to discriminate P1B-ATPases in different organisms based on their structural amino acid features. In addition, for the first time, reliable models for prediction of the hyperaccumulating activity of unknown P1B-ATPase pumps were developed.

We employed our method in monitoring and prediction of breast cancer. The results confirmed that amino acid composition can be used to discriminate between proteins groups expressed in two forms of breast cancer: malignant and benign. This study was strong evidence that malignancy can be predicted out from amino acid, and malignant proteins can be distinguished based on the amino acid composition of their proteomes without further need to protein separation. An important outcome was discovery the role of dipeptides, in particular Ile-Ile, in cancer progression. In addition, Generalized Rule Induction (GRI) found association rules found in the data showing 100 most important rules classifying benign, malignant, and common expressed proteins expressed in breast cancers.

In another investigation, we found that EST-SSRs in normal lung tissues are different than in unhealthy tissues, and tagged ESTs with SSRs cause remarkable differences in amino acid and protein expression patterns in cancerous tissue. This can be supposed as a glimpse of invention of new sort of biomarkers based on frequency of amino acids.

Up to now, phylogenic trees, drawn by nucleic acid or amino acid sequence alignments, have employed as the base of evolutionary studies. However, this method does not take into account the structural and functional features of sequences during evolution. On the contrary, the presented classification here, based on the decision

tree, anomaly detection model and feature weighting, provides an evolutionary separation of organisms based on their structural reasons of this diversity.

Our findings have the potential to be efficiently used in the following area: filling the gap between laboratory engineering of proteins and computational biology, developing amino acid feature based-biomarkers, increasing the accuracy of prediction of 3D protein structure based on important amino acid features, and developing websites/softwares for prediction of the result of mutation. In addition, important discovered amino acid features can be employed as clues for discovering important DNA mutations and increasing prediction accuracy of 3D structure from DNA sequence. Furthermore, this study offers new for protein function, irrespective of similarity searches.

Background

To make a protein, DNA/RNA (nucleic acid) translates to amino acid sequence, subsequently; amino acids go through various folding and build primary, secondary, tertiary, quaternary, and in final, 3D structure of protein to perform its function. Regarding the central role of protein in diseases, cancer, pathogen mutation, and any other biological phenomena, one of the major challenges in science is to determine protein structure.

Experimental determination of protein structure

X-ray crystallography and NMR spectroscopy are experimental methods in determining protein structure, but both methods are expensive, time-consuming, and cannot be applied for all proteins. In fact, due to commonly a large number of proteins involved in a biological event, it is practically impossible to obtain the experimentally protein images of all proteins.

Prediction of protein structure and function

Enormous computational efforts have been carried out to predict protein structure and function. However, nearly all of these efforts have been focused on prediction of function based on primary nucleic acid sequence or modelling 3D structure of protein from its nucleic acid sequence. In fact, it seems that amino acid attributes, which is an intermediate phase between DNA/RNA and advanced protein structure, have been missed.

Our theory

For a long time, the general belief was that amino acid features, such as frequency of amino acids, are primary features with restricted number which can not be evaluated for understanding complex structural protein modulaton in events such as thermostability, halostability, or cancer. In fact, it was thought that DNA/RNA produces more information than amino acids, and 3D structure is near to real protein figure in nature.

From 2010, we examined the possibility of precise prediction of structural protein function based on amino acid features by improving the following three aspects of amino acid research:

- ✓ Increasing the number of calculated amino acid features
- ✓ Testing different feature selection (attribute weighting) algorithms and selection of the most important amino acid attribute based on the overall conclusion of algorithms
- ✓ Examining different supervised and unsupervised data mining (machine learning) algorithms
- ✓ Joining attribute weighting with different data mining algorithms
- ✓ Applying the procedure in different biological examples including: thermostability, halostability, prediction of function of heavy metal transporters, cancer prediction, and pursuing the EST-SSRs in amino acid and protein level

Increasing the number of amino acid features - importance of dipeptides

We suggest that the major drawback of previous attempts was considering a small number (less than 50 amino acid features), regarding the fact that many attributes determine the different characteristics of a protein molecule. Interestingly, by developing of different bioinformatics tools, it is possible to calculate a large number of amino acid characteristics from each amino acid sequence. At first, we increased the list of computed amino acid attributes from primary amino acid sequence to more than 800 features [1-6]. In addition to frequency/count of different amino acids, frequency of hydrophilic/hydrophobic residues, Count/frequency of negatively/positively charged, aliphatic index, count and frequency of H, C, N, O, and S, count/percent of alpha helix, count/percent of Pi helix, count/percent of extended

strand, count/percent of beta turn, etc, a particular attention was paid to the role of dipepides. Surprisingly, up to now, there has been little discussion about the role of dipeptides in protein function. Our studies demonstrated that specific dipeptides, such as Asn–Gln, play the central role in protein halostability [5] and discrimination of halotolerant and halo-sensitive proteins. Furthermore, critical role of dipeptides, such as Asp-Gln, were highlighted in thermostability of protein [4].

Attribute weighting

Due to increase in a number of features and possible complexity in analysing the results, we examined different feature selection (attribute weighting) algorithms to address this concern. The results showed that in all studies, attribute weighting algorithms can be effectively be used in finding the important features[1-6]. Attribute weighting models create a more manageable set of attributes for modeling by reducing the size of attributes. In our studies, we used 10 different attribute weighting algorithms including weighting by PCA, SVM, Relief, Uncertainty, Gini index, Chi Squared, Deviation, Rule, Correlation, and Information Gain. Each attribute weighting system uses a specific pattern to define the most important features. Thus, the results may be different. Our strategy was to select the important features was simple but efficient; features which have been announced important by the majority of attribute weighting methods were selected as important.

Employed data mining procedures

We found that data mining (machine learning) models have the great potential to link amino acid features of protein structure. While commonly, researchers select a small number of machine learning algorithms (based on previous studies, etc), one of the strengths of our approach was examining different supervised and unsupervised data mining algorithms with/without application of preattribute weighting methods to find the best model to fit the biologicl data.

Unsupervised clustering algorithms

To organize data into a more meaningful form, based on aminoacid fetures, clustering algorithms partition the data into groups or clusters according to various criteria. Four different unsupervised clustering algorithms including, K-Means, K- Medoids, Support Vector Clustering (SVC), and Expectation Maximization Clustering (EMC)

were applied on ten datasets created using attribute weighting as well as original data set (in total, 11 datasets). A suitable unsupervised algorithm was capable of discovering structure on its own by exploring similarities or differences between individual data points in a dataset.

Supervised Clustering

We applied different neural network and decision tree supervised clustering methods in our studies[1,4-6].

Decision Trees. Six tree induction models including Decision Tree, Decision Tree Parallel, Decision Stump, Random Tree, ID3 Numerical, and Random Forest were run on the main dataset as well as 10 pre-treated datasets with attribute weighing algorithms. Each tree induction model ran with the following four different criteria: Gain Ratio, Information Gain, Gini Index and Accuracy. In addition, a weight-based parallel decision tree model, which learns a pruned decision tree based on an arbitrary feature relevance test (attribute weighting scheme as inner operator), was run with 13 different weighting criteria (SVM, Gini Index, Uncertainty, PCA, Chi Squared, Rule, Relief, Information Gain, Information Gain Ratio, Deviation, Correlation, Value Average, and Tree Importance)[4]. In other words, for examples such as thermostability, 37 different Decision Tree algorithms were tested for each condition of dataset (one original and 10-treated with attribute weighting).

Neural Networks. Artificial neural networks are computing systems that simulate the biological neural systems of a human brain. We run Feed-forward and Elman (as a type of a recurrent) neural networks with various hidden layers in each neural network [4]. The first one was original dataset and the next one was dataset after application of stepwise feature selection algorithm. The stepwise regression feature selection algorithm was applied to identify the attributes that had a strong correlation with target variable (protein function) such as thermostability [4].

Cross validation

Cross validation was used to evaluate the accuracy of models by training and testing. Data in each feature dataset was divided into 10 nearly equal classes, and the accuracy was calculated from the number of correct determined records by the model divided by all the records in each class. The process was repeated 10 times and the accuracy for true, false and total accuracy calculated. The final accuracy was the average of the

accuracy in all 10 tests[1,4,6].

In addition, in examples such as thermostability[4], the efficiency of predicted neural networks were tested with new protein dataset with known thermostability which has not been used in training and testing. The calculated accuracy indicated the performance of the developed models in predicting the thermostability of new sequences.

Example of applications of the new method

Thermostability

A small portion of enzymes can withstand higher temperatures as a result of various structural adaptations. Understanding the protein attributes that are involved in this adaptation is the first step toward engineering thermostable enzymes. The aim of our study here was to determine the most important amino acid attributes that contribute to protein thermostability. As mentioned above, we examined a variety of attribute weighting algorithms and various supervised and unsupervised clustering models on a large number (800) of amino acid properties. Seventy percent of the weighting algorithms selected Gln content, frequency of Asn-Glu dipeptide, and frequency of hydrophilic residues as the most important protein attributes to distinguish between mesostable and thermostable enzymes.

More importantly, we successfully established an accurate expert system to predict the thermostability of any input sequence trough mining of its calculated amino acid features. Interestingly, performance of a clustering algorithm such as EMC can vary from 0.0% to 100%, depending upon which attribute weighting algorithm had summarized the attributes (features) of the dataset prior to running the clustering algorithm. The highest accuracy (100%) was observed when the EMC clustering method was applied to datasets generated by Correlation and Uncertainty attribute weighting algorithms[4]. This finding highlights the importance of testing different attribute weighting algorithms in biological studies; particular attribute weighting may be unique to each biological case.

Halostability

Recently, in another attempt, we analyzed the performance of different attribute weighting, screening, clustering, and decision tree algorithms to discriminate

halophilic and non-halophilic proteins[4]. The results showed that amino acid composition can be used to efficiently discriminate halostable protein groups with up to 98% accuracy. Similar to thermostability, these results showe precise prediction of halostability when an appropriate machine learning algorithm mines a large number of structural amino acid attributes of primary protein structure[4].

Phytoremediating proteins

Phytoremediation is the use of plants for extraction and detoxification of pollutants which provides a new and powerful approach against a polluted environment. Small numbers of plants, such as Thlaspi spp, take advantages of highly efficient heavy metal ATPases in overall metal ion homeostasis and hyperaccumulation. P1B-ATPases pump a wide range of cations, especially heavy metals, across membranes against their electrochemical gradients.

In another recent study, we applied diverse weighting and modeling approaches to build a precise model to discriminate P1B-ATPase heavy metal transporters in different organisms based on their structural protein features[1]. We analyzed 2644 protein characteristics of primary, secondary, and tertiary structures of P1B-ATPases in hyperaccumulator and nonhyperaccumulator plants to identify differences between proteins in hyperaccumulator and nonhyperaccumulator pumps.

Surprisely, similar to thermostability and halostability, simple amino acid features, without the need of advanced features of protein structure, could explain the difference between P1B-ATPases in hyperaccumulator and nonhyperaccumulator plants. Glycine count, frequency of glutaminevaline, and count of valine-phenylalanine were the most important attributes highlighted by 10, 5, and 4 weighting and decision tree models, respectively. More importantly, a precise model was built to discriminate P1B-ATPases in different organisms based on their structural amino acid features. In addition, reliable models for prediction of the hyperaccumulating activity of unknown P1B-ATPase pumps were developed[1]. Uncovering important structural features of hyperaccumulator pumps in this study provided the knowledge required for future modification and engineering of these pumps by techniques such as site-directed mutagenesis. In this study, decision trees showed high performance.

Amino acid based-decision trees: a new avenue in phylogenic analysis and unravelling the underlying of outstanding proteins

A decision tree is constructed by looking for regularities in data, determining the features to add at the next level of the tree using an entropy calculation, and then choosing the feature that minimizes the entropy impurity [7].

It seems that application of decision trees in protein science can offer various benefits including:

- ✓ Discovering important attributes, and more importantly visualized ranking of attributes based on their importance. In contrast, attribute weighting algorithms are able to show the important features, but they can not rank them and discover the routes
- ✓ Informative prediction. The superiority of decision tree algorithms over neural networks is that prediction is visualized and underlying background of prediction can be understood clearly. In contrast, this opportunity is not provided in neural networks.

In addition, we suggest that designed decision trees, based on amino acid features, open a new vista in phylogenetic analysis that we call that smart phylogenetic trees strategy. In fact, it can be assumed that the more variable amino acid features in primary branches of the decision trees are key amino acid features for evolution of proteins. The major benefit of decision tree based-phylogenetic analysis of amino acid attributes is that clustering is accompanied with the structural (functional) information of protein packaging. In addition, main branches of decision tree determine the most important features in structural packaging of the protein; and sub-branches are built based on features with lower rank of importance. Interestingly, new branches of phylogenetic tree can be supposed as different possible routes of evolution to induce the protein.

As example, we applied this novel strategy of phylogenic analysis in comparative study of ammonium transporters in different organisms. Ammonium is an excellent nitrogen source, and ammonium transfer is a fundamental process in most organisms. Membrane transport of ammonium is the key component of nitrogen metabolism mediated by ammonium transporter/methylamine permease/rhesus (AMT/MEP/Rh) protein family. We created an accurate and precise links between amino acid and ammonium transporters in different organisms (animal, fungi, plant, bacteria, and

human) by extracting and calculating of a large number of structural amino acid characteristics by various weighting and modelling algorithms[8].

The results, for the first time, indicated that His-based features including count/frequency of His and frequency/count of Ile-His were the most significant features generating different types of ammonium transporters within organisms. Within different tested models, the C5.0 decision tree model was the most efficient and precise model for discrimination of organism type, based on ammonium transporter sequence, with the precision of 94.85%.

Determination of protein characteristics of ammonium transporters in different organisms provides a new vista for understanding the evolution of transporters based on the modulation of amino acid features. In addition to providing a basis for future studies on engineering novel ammonium transporters, dissecting a large number of structural protein characteristics through decision tree algorithms provides a novel functional strategy for studying evolution and phylogeny.

Up to now, phylogenic trees, drawn by nucleic acid or amino acid sequence alignments, have employed as the base of evolutionary studies. However, this method does not take into account the structural and functional features of sequences during evolution. On the contrary, our presented classification, based on the decision tree, anomaly detection model and feature weighting, provides an evolutionary separation of organisms based on their structural reasons of this diversity. The presented procedure can highly enrich and qualify any type of further evolutionary study by the completion of common phylogenic analysis methods.

Amino acid-based monitoring of cancer progression: a new avenue in cancer discovery and prediction

Early diagnosis of breast cancer is much more significant than any treatment; therefore, more attention has been paid to discover the methods for early diagnosis of cancer. By developing new technologies in sequencing and possibility of rapid full genome sequencing of patients as well as rapid growth in 2-dimentional-based protean separation techniques, most studies have concentrated on prediction based on DNA mutation or proteome alteration. Consequently, most studies are seeking for DNA or protein biomarker discovery.

In our studies for the first time, we examined the idea of following the intermediate amino acid alteration of different cancers. We were interested to know whether cancer progression is accompanied with considerable shift in structural amino acid features.

In primary attempts, we analysed breast cancer and lung cancer progression[2,3]. In breast cancer study, at first, more than 800 features of proteins expressed in malignant, benign, and both cancers were compared using different screening, clustering, decision tree, and generalized rule induction (GRI) algorithms to look for amino acid patterns in two benign and malignant breast cancer groups with and without running of feature selection[2].

The results showed that 57 out of 838 amino acid features rank as important ($p > 0.05$). The depth of the trees induced by tree induction models varied from 5 (in the Quest model) to 2 (in the C5.0 model) branches. The accuracy was between 86-100%. The best performance evaluation found when C&RT model was applied. Interestingly, the frequency of a di-peptide, Ile-Ile, was the most important protein attributes in all tree and rule induction models.

Generalized rule induction (GRI) found association rules found in the data showing 100 most important rules classifying benign, malignant, and common expressed proteins expressed in breast cancers. In all tree induction models, the count of Ile-Ile chose as the most important attribute and also in all GRI association rules (100 rules) the count of this feature was used as an antecedent to support the rules[2].

The results confirmed that amino acid composition can be used to discriminate between proteins groups expressed in two forms of breast cancer. This study is strong evidence that malignancy can be predicted out from amino acid, and malignant proteins can be distinguished based on the amino acid composition of their proteomes without further need to protein separation. An important outcome was discovery the role of dipeptides, in particular Ile-Ile, in cancer progression.

In an interesting recent study, we found that during lung cancer induction, some SSRs, in particular trinucleotides, intentionally join specific ESTs in cancerous tissue such as chromodomain helicase protein.

Interestingly, we observed that similar to the EST level, the expression pattern of EST-SSRs-derived amino acids was significantly different between normal and cancerous tissues[3]. Arg, Pro, Ser, Gly, and Lys were the most abundant amino acids in cancerous tissues, and Leu, Cys, Phe, and His were significantly more abundant in normal tissues. For the first time, this study confirmed that EST-SSRs in normal lung

tissues are different than in unhealthy tissues, and tagged ESTs with SSRs cause remarkable differences in amino acid and protein expression patterns in cancerous tissue. This can be supposed as a glimpse of invention of new sort of biomarkers based on frequency of amino acids.

Application of amino acid features in determining of influenza virus A subtype and predction of new hosts

Influenza A viruses infect large numbers of animals and are subtyped according to their surface antigens to 16 HA subtypes and 9 NA subtypes. In recent study, to identify the main prominent protein attributes representing each subtype, we applied various clustering, screening, item set mining and decision tree models applied to dataset of 3632 HA sequences of influenza A viruses. Nine hundred and twenty four amino acid features were computationally calculated for each HA sequences[9].

The count of Tyr, Gln, Phe, and the count of specific hydrophilic – hydrophobic dipeptides (such as Lys – Val, Asn – Leu and Pro – Leu) were the most important protein features. Most decision tree models used non-reduced absorption at 280nm as the main protein feature to build the trees. Parallel stump and ID3 numeric decision tree algorithms were the best tree to differentiate between HA subtypes. For the first time, this paper showed that protein attributes can be used to differentiate between influenza A subtypes.

Application of this discovery

Filling the gap between laboratory engineering of proteins and computational biology

The amino acid sequence (primary structure) of a protein is the main indicator of its function. However, it is generally agreed that direct prediction of protein characteristics such as thermostability, halostability, or high-affinity transporter from the primary amino acid sequence is not possible. Up to know, computational methods to predict protein function have focused on tertiary and quaternary structures (i.e., threedimensional structure and molecular protein volume). Consequently, further advances have been hindered by the difficulties in manipulating these complex features. In fact, there is a gap between computational biology results and laboratory

applications, because available laboratory techniques are limited to simple substitution of a small number of amino acids in the primary protein structure.

Amino acid feature based biomarkers

The primary results of application of this method in lung and breast cancers[2,3] offer the new source of markers based on amino acid features. In particular, it seems that dipeptides, which have not been considered widely before, are a good target in this field.

Increasing the accuracy of prediction of 3D protein structure based on important amino acid features

Findings of this research have the potential to modify/improve the 3D structural protein prediction from DNA/RNA sequence. In fact, important-announced amino acid features and more importantly, their importance ranking based on algorithm such as decision tree, can be used as a clue for functional improvement of available computational modelling methods of predicting 3D structure from nucleotide acid sequence.

Important discovered amino acid features: a clue for discovering important DNA mutations

Rapid development of sequencing project has generated massive amount of DNA sequences and DNA mutations. The major question has been raised here that which of these mutations are important in induction of disease and caners?.

We believe that our procedure by discovering important amino acid features in cancer and disease by data mining algorithms provides smart strategy for selection of key DNA mutations. Important announced structural amino acid features can be transferred to DNA level and discovered key DNA mutations. High accuracy of the found models in different biological examples (higher than 90%) can guaranty the success of this method.

Important discovered amino acid features: a clue for more precise prediction of 3D structure from DNA sequence

There is a great interest on developing precise 3D structure of protein from DNA sequence. Protein folding algorithms are based on finding the global minimum of the

free energy of the polypeptide chain by physical simulations or by a guided search in conformational space usingempirical molecular potentials[10]. The presented method in this study on finding important amino acid features in the function of protein provides a new index and can be efficiently used in improving the accuracy of prediction of 3D structure from DNA sequence. In fact, in selection of the best possible status of predicted 3D structure, in addition to indices such as optimum low free energy, corresponding with the functionally announced amino acid features can be considered.

Novel strategy for protein function, irrespective of similarity searches

The major achievement of this discovery was the prediction of the thermostability/halostability/activity of heavey metal trasporter/Influenza virulent protein etc from any input protein sequence based on its amino acid composition. This study demonstrates the feasibility of predicting protein function irrespective of sequence similarity and will serve as a basis for engineering of enzymes. In the present method, there is no need to utilise any sequence similarity searches or protein tertiary/quaternary features

We can have an estimation of the function of any generated or mutated sequence by calculating the amino acid properties of that sequence and running the previously obtained data mining models.

Developing websites/softwares for prediction of the result of mutation

The mutation has very low success frequency regarding the fact that more than 98% of mutations are destructive. However, using the procedure described in this study, it is possible to evaluate the function of any possible mutations by models resulting huge decrease in laboratory works.

These predictions did not require similarity searches or gathering information about the complex, expensive, and time-consuming features of the tertiary and quaternary protein structure.

The developed models can be further embedded in web-based data banks or softwares to predict protein function of possible mutations before laboratory examination.

New strategy for phylogenetic analysis

Up to now, phylogenic trees, drawn by nucleic acid or amino acid sequence alignments, have employed as the base of evolutionary studies. However, this method does not take into account the structural and functional features of sequences during evolution. On the contrary, the presented classification here, based on the decision tree, anomaly detection model and feature weighting, provides an evolutionary separation of organisms based on their structural reasons of this diversity.

Future works

Cancer studies

In the frame of Discovery Project of ARC (Australian Research Council), we tend to examine this method in different cancers such as leukaemia, ovarian cancer, etc, to develop prediction methods and also improve DNA mutation discovery and 3 D structure base prediction of altered protein.

In addition, we would to continue the work on finding association rules and dipeptides in different cancer since this part of study provides valuable information in cancer studies.

Linking important discovered amino acids with transcription factor

In the above mentioned project, we tend to link between amino acid and regulatory network of gene activation.

Influenza Research

In the frame of another Discovery project, by employing our approach, we are keen to find that which amino acid features produce this ability for influenza virus to increase the range of its hosts.

Moreover, by developing amino acid based prediction models, we intend to predict new strains of virus.

1 Ashrafi, E., Alemzadeh, A., Ebrahimi, M., Ebrahimie, E. & Dadkhodaei, N. Amino Acid Features of P1B-ATPase Heavy Metal Transporters Enabling Small Numbers of Organisms to Cope with Heavy Metal Pollution. *Bioinform Biol Insights* **5**, 59-82, doi:10.4137/BBI.S6206 bbi-2011-059 [pii] (2011).

2 Ebrahimi, M., Ebrahimie, E. & Shamabadi, N. Are there any differences between features of proteins expressed in malignant and benign breast cancers? *J Res Med Sci* **15**, 299-309 (2010).

3 Bakhtiarizadeh, M. R., Ebrahimi, M. & Ebrahimie, E. Discovery of EST-SSRs in Lung Cancer: Tagged ESTs with SSRs Lead to Differential Amino Acid and Protein Expression Patterns in Cancerous Tissues. *PLoS One* **6**, e27118, doi:10.1371/journal.pone.0027118
PONE-D-11-13096 [pii] (2011).

4 Ebrahimi, M., Lakizadeh, A., Agha-Golzadeh, P. & Ebrahimie, E. Prediction of thermostability from amino acid attributes by combination of clustering with attribute weighting: a new vista in engineering enzymes. *PLoS One* **6**, e23146, doi:10.1371/journal.pone.0023146
PONE-D-11-06772 [pii] (2011).

5 Ebrahimie, E., Ebrahimi, M. & Sarvestani, N. R. Protein attributes contribute to halo-stability, bioinformatics approach. *Saline Systems* **7**, 1, doi:1746-1448-7-1 [pii]
10.1186/1746-1448-7-1 (2011).

6 Ebrahimi, M. & Ebrahimie, E. Sequence-Based Prediction of Enzyme Thermostability Through Bioinformatics Algorithms. *Curr Bioinform* **5**, 195-203 (2010).

7 Huang, L.-T., Gromiha, M. M. & Ho, S.-Y. iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations. *Bioinformatics* **23**, 1292-1293, doi:10.1093/bioinformatics/btm100 (2007).

8 Tahrokh, E. *et al.* Comparative study of ammonium transporters in different organisms by study of a large number of structural protein features via data mining algorithms. *Genes & Genomics* **33**, 565-575, doi:10.1007/s13258-011-0057-6 (2011).

9 Ebrahimi, M., Agha-Golzadeh, P., Ebrahimie, E. & Shamabadi, N. Application of bioinformatics models to define influenza virus A subtypes. *Proceedings of the 2011 International Conference on Bioinformatics & Computational Biology* 513 - 515 (2011).

10 Marks, D. S. *et al.* Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLoS One* **6**, e28766, doi:10.1371/journal.pone.0028766 (2011).