Protein Classification using Hidden Markov Models and Randomised Decision Trees

Arron Lacey, Jingjing Deng, and Xianghua Xie
Department of Computer Science, Swansea University, Swansea SA2 8PP, United Kingdom
http://csvision.swan.ac.uk

***Abstract*—Since the introduction of next generation sequencing there is a demand for sophisticated methods to classify proteins based on sequence data. Two main approaches for this task are to use the raw sequence data and align them against other sequences, or to extract discrete high level features from the protein sequences and compare the features. Two machine learning methods are demonstrated to show each approach. Profile Hidden Markov Models are built from multiple alignment of raw sequence data and learn amino acid emission and transition parameters for a given alignment and effectively harness the power of aligning a test protein to a model built form many proteins. Random Forests on the other hand are used to discriminate between two sets of proteins based on features such as functional amino acid groups and physiochemical properties extracted from the raw sequences. The strengths and limitations of each method are presented and discussed, focussing on the individual merits and how they could work possibly compliment each other rather than just being compared by their classification accuracy.**

I. Introduction

The problem of protein family classification in biology is one that has benefited greatly from the application of machine learning and pattern recognition techniques. Research teams world-wide curate electronic biological databases of proteins sequenced from organisms, of which the size of such databases increase exponentially, and it is the role of machine learning and automated pattern recognition techniques to ensure that the function and structure of these proteins are analysed at the same rate proteins sequences are made available in the public domain. In biological terms, two proteins may be related based on common patterns found in the sequences, where families of proteins are typically classified by the functional purpose. It is therefore not only interesting to know which family a protein belongs to, but also what features in the sequence are common to those within the family. The two main approaches for classification are to (i) to use the raw sequences and align them, where in alignment space common sub-strings are identified and scored based on metrics such as whether a certain sub-string is conserved in nature, or (ii) use high dimensional meta data extracted from the sequence such as hydrophilic scale [1] where the features are arbitrarily pre-defined.

In the literature it has often been the case that regardless of whether the technique uses meta data or actual sequence data, it is usually the classification accuracy that is compared rather than the merits of each technique, even though many works show similar accuracy.

The statistics behind position-specific scoring based methods on pairwise alignments of proteins have been established in [2] and their work on these scoring systems produced BLAST (Basic Local Alignment Search Tool), where sub strings of two sequences are compared when aligning them, and each substring in the alignment is scored based on matches and mis-matches between the two.

```
VSPAGMASGYD
:  |  |  ||  ||
I-P-GKAS-YD
```

Fig. 1. A simple pairwise alignment. The alignment shows the two may be related based on matched amino acids in the sequence, where some amino acid substitutions are tolerated in nature better than others. It can also be seen in the sequence that some amino acid deletions may have occurred over time and dashes represent these when aligned to the first sequence.

Fig. 1 provides an example alignment which shows how it may be possible to related one sequence to another for pairwise alignments, however a more sophisticated scoring system is desired that could harness the power of aligning multiple sequences. Multiple sequence models were introduced by Taylor *et al.* in 1986 [3] and further developed by Henikoff *et al.* [4] and Eddy [5] as a means to use position-specific information from *defined* sequence alignments. Traditional HMM (Hidden Markov Model) provide a method to determine what state a system is in based on emitted symbols, such as those from a protein sequence. Profile HMM builds upon this by determining how likely a symbol in a sequence is emitted in a certain position of a multiple sequence alignment, as well as modelling the probability of transitioning to an insert or delete state. HMM can be built on either aligned or unaligned sequences, where a previous multiple sequence alignment may used if the inserts and deletion in a protein may be of interest.

From the multiple sequence alignment, the probability of all match, insert and delete states at each position in the multiple sequence alignment are determined through some training methods, such as the forward-backward or Viterbi methods. New sequences can then be aligned to the model and scored based on the path it takes through the HMM model. The success of profile HMM has been shown by Eddy [5] where the PFAM database [6] holds information on protein family domains built entirely from profile HMM.

In contrast to algorithms such as BLAST and HMM that

978-1-4799-5838-2/14/$31.00 ©2014 IEEE

use alignments of raw sequences to classify proteins, there are many other techniques such as artificial neural networks, SVM (Support Vector Machine) and RDT (Random Decision Tree) that use meta data extracted from raw sequence data. Statistics ranging from simple frequencies of amino acids to functional groups, secondary protein structure all frequently used as pre-defined inputs to such classifiers. Randomized decision trees as classifiers have shown accurate results in protein classification in the literature, although not nearly as widely used as popular techniques such as SVM and ANN (Artificial Neural Network). RF (Random Forests) is an ensemble machine learning technique which builds decision trees at training time to output classes within the training set based on splitting the data at each node by a threshold. For each tree in the forest, the tree is trained and tested using bootstrapped samples (with replacement) of the dataset where the test data is referred to as OOB ("Out of Bag") data that is used to estimate an OOB error. This is particularly useful to biologists trying to classify proteins, as it allows training and testing to be tailored towards certain proteins features, for example groups of amino acids that represent hydrophilic in a protein.

Kandaswamy *et al.* [7] demonstrated RF to be a successful classifier for antifreeze proteins when using non-antifreeze proteins as a negative test set, achieving 84% accuracy, which are better than other methods used such as HMM, SVM and ANN in their study. RF can be used and should be explored for other protein families to be established as a tool for future classification when new proteins are found. Another useful feature from the RF algorithm that has been explored in the literature is feature importance using measures such as Gini importance and permutative importance. Feature importance is an integral part of protein-protein interaction studies as it explains the relationships between a protein bonds, and as this experiment shows, the Gini importance picks out features known in the literature to be essential features as part of tansmembrane and antifreeze proteins that are used to split the classify the proteins best over a range of other features. The work in this paper describes the implementation of hidden Markov models and random forests for protein classification and the strengths and weaknesses of both when analysing different groups of proteins.

II. Dataset

The following two experiments use the RF and HMM to classify two different types of protein families: ion channel transmembrane proteins from non transmembrane proteins and antifreeze from antifreeze-like proteins. Transmembrane proteins exist within the membranes of cells that transport molecules and ions across the membrane to inside the cell. Transmembrane proteins show high structural homology across the family. In contrast, antifreeze proteins do not show high structural homology, however are generally arranged in such a way that water molecules do not unfold them. The type III clan consists of two sub groups: one being antifreeze and similar proteins such as flagella and pilus proteins that provide a similar functional role, and the second being homologous proteins in terms of function, which for ease of use will be referred to as antifreeze and antifreeze-like proteins respectively. This particular family of antifreeze proteins have been chosen in contrast to the transmembrane proteins. Antifreeze proteins do not have such well defined structure because they have conversantly evolved from various different types of organisms [8] and as such the high variance in structure of the subtype constituents if each family will be a good test for classification.

A. HMM data

The PFAM database [9] stores protein family data built using HMM. Three sub-types of transmembrane proteins were obtained: 3732 ligand-gated ion channel (*PF00060* 44 training, 3228 testing sequences), potassium-transporting ATPase A subunit proteins (*PF03814* 14 training, 2239 testing sequences) and inward rectifier potassium channel (*PF01007* 14 training, 1452 testing sequences). A negative training dataset consisting of 1445 randomly selected non-transmembrane proteins were also obtained from PFAM. For the antifreeze proteins, type III antifreeze proteins were obtained from the PFAM database, where the family is split into antifreeze proteins (*PF086666* 169 training, 4935 testing sequences) and their homologous antifreeze-like proteins (*PF13144* 119 training, 1927 testing sequences). All proteins used to train and test the HMM were pre-aligned to include insertions and deletions in the sequences.

B. Random Forests data

Meta data extracted from raw sequence data can take a long time depending on what features are desired to aid classificaton and as such smaller test sets were used in the random forests data. 337 voltage gated ion channel transmembrane proteins (297 training, 40 test sequences) were taken from the Transporter Classification Database [10] and 297 and 40 non transmembrane respective training and testing sets were taken from PFAM. A training set of 100 antifreeze proteins (*PF086666*) and 100 antifreeze-like proteins (*PF13144*) were taken from the PFAM database, and 26 antifreeze proteins and 26 antifreeze-like proteins were used as test sets.

III. Model and Feature

Profile hidden Markov models are available from PFAM website, or can be built using the raw sequences contained within the family. In this paper, in total 5 HMMs were built (3 transmembrane and 2 antifreeze), in which the length after aligning all training set proteins was used to generate each model and. was taken as the model length. The Baum-Welch learning algorithm is used to estimate the transition and emission matrices. Fig. 2 illustrates the HMM model. Once each model is built, test proteins can be aligned to the model using the Viterbi algorithm and scored against that HMM model. Three classification tests were devised to show the strengths and weaknesses of profile HMM: i) Transmembrane vs. non-Transmembrane proteins ii) Ligand transmembrane

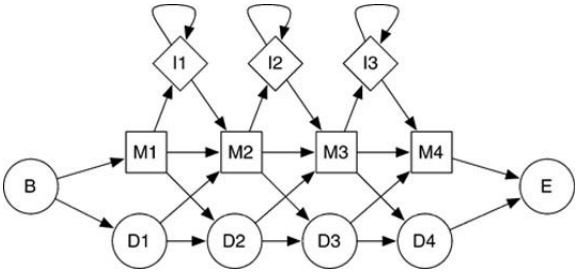

Fig. 2. HMM model. There are three states of a profile HMM: match,insert and delete states. Match states emit the amino acid observed at that position of the sequence. Insert states occur when there is is an amino acid(s) inserted in the sequence which also emit 1 of 20 different amino acids, which can be pre-configured to emit background amino acids or amino acids known to be found in that particular protein. Delete states are silent states that emit no amino acid.

proteins vs. potassium and inward rectifier transmembrane proteins, and iii) antifreeze vs. antifreeze-like proteins. High level meta-data was extracted from the random forest data using a variety of techniques and are summarised as follows:

Protein features

#	**Features**	No. of features
1	Amino acids	20
2	Functional groups	17
3	Chemical properties	6
4	Secondary Structure	60
	Total	**93**

1) *Frequency of amino acids:* The frequency of each of the 20 naturally occuring acids was calculated.
2) *Frequency of functional groups:* The amino acids found within each protein sequence were categorized into 17 functional groups such as phenylene, valine, leucine, proline and hydroxyl, where the frequency of each functional group was calculated.
3) *Secondary Structure* Frequency of helix, beta sheet and coil structures within each protein were predicted using PSIPRED. The frequency of each amino acid found within each secondary structure element was calculated. PSIPRED [11].
4) *Physio-chemical properties:* The frequency of physio-chemical properties of each protein was derived from the amino acid index (AAINDEX) [12] database, and their methodologies were used to calculate the isoelectric point, aromaticity, grand average of hydropathicity index, instability index as well as molecular mass of all protein sequences [13].

IV. Feature Importance

The measure used in this work to determine which features best split the data from the RF algorithm is the Gini index. The Gini index is essentially measured by calculating the level of impurity of the data at each node split found within the child nodes. At each node j, the impurity, or "Gini impurity" $G(j)$ is defined as:

$$G(j) = 1 - p_1^2 - p_2^2 \tag{1}$$

where $p_k = n_k/n$ is the fraction of n_k samples from class $k = 0, 1$ from n samples at the node j. The change in the Gini impurity as the data is split into two child nodes is

$$\delta G(j) = G(j) - p_L G(j_L) - p_R G(j_R) \tag{2}$$

where p_L and p_R are the respective sample fractions held in the child nodes. At each node, an exhaustive search over features and thresholds yields a pair Φ, τ that represents the maxmimum value of $G(j)$ that decreases with each node split, and for each node j in each tree T the Gini importance is the sum of all pairs yielding maximum $G(j)$

$$I_G(\Phi) = \sum_T \sum_j \delta G(j)_\Phi(j, T) \tag{3}$$

$I_G(\Phi)$ is a measure of how often feature Φ was used to split a node. If the Gini index decreases at each node, then clearly the larger the Gini importance for Φ, the more important that feature is in classifying the data.

Where accurate classification is a prequisite for any protein classifaction experiment, feature importance is perhaps more interesting than simply comparing different algorithms to see which classifier performs best as it is important to understand the nuances of each protein in biological terms i.e why they belong to that family. It has to be noted however that any classification and feature extraction of random forests are only relative to the training sets used. A protein is suspected to belong to a family then the positive dataset is not so much of a problem, but the negative dataset equally important. Therefore negative training sets should be devised and randomly selected to represent background frequencies of any of the features used as input into the training process.

A comparison of the feature importance between the frequency of amino acids and the entire feature set in the transmembrane proteins is shown in Fig. 3. The frequency of phenylene denoted by FH (Helix Positions), closely followed by molecular weight were found to be the two most important features for classifying the dataset as measured by the Gini index.

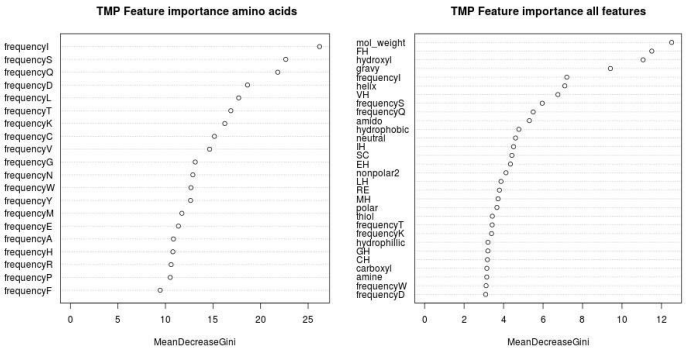

Fig. 3. Feature importance of frequency of amino acids against the entire feature set in transmembrane proteins.

Phenylene helices are integral in promoting folding of proteins to perform the functions that transmembrane proteins carry out in cells, and are also involved initiating interaction between other transmembrane proteins [14], and so confirms the validity of using the Gini Index for feature importance measures. The role of molecular weight in any type of analysis between proteins is trivial in terms of function, yet useful for discriminatory measures. It is encouraging that the random forest process used the GRAVY to split the data. Kyte and Doolittle's work on hydropathy [1] showed that transmembrane proteins will have a higher GRAVY score than other globular proteins. Other notable features of the data that would perhaps be expected to help determine between a transmembrane protein and a non-transmembrane protein that are also found to be of importance in this study are hydroxyl groups typically found in the form of glycerol that are found in cellular membranes, and the frequency of amino acids found in helices of each protein. Helices are responsible for the structure of transmembrane proteins where Bowie *et al* [15] documented"helix packing" in transmembrane proteins, but will fall to background frequencies in globular proteins.

A comparison of the feature importance between the frequency of amino acids and the entire feature set in the antifreeze proteins is shown in Fig. 4. Molecular weight being the most important feature to classify the data does not really have much meaning, in particular from the biological function point of view, as on average the non homologous antifreeze-like proteins have even longer sequence than the antifreeze proteins. Glutamate (Q), pointed out as an important feature has been proven to be an essential solute that increases the ability for antifreeze proteins to increase thermal hysteris four-fold [16] The absence of functional groups in data splitting between the antifreeze proteins and antifreeze-like proteins would be expected as functionally they are near identical while being structurally different, as seen by the amount of secondary structure features splitting the data.

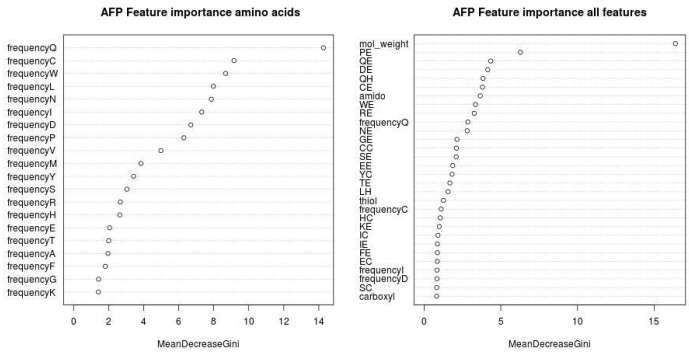

Fig. 4. Feature importance of frequency of amino acids against the entire feature set in antifreeze proteins.

Although random forest classification of antifreeze and non-antifreeze proteins have been reported by Kandaswamy *et al.* at just under 84% accuracy, a dataset of antifreeze and antifreeze-like proteins were chosen to test the discriminative power of random forests. It would be expected that it would be difficult to classify two sub families of proteins rather than classifying antifreeze from non-antifreeze proteins. On the face of the results in this experiment a baseline accuracy of 86% was achieved leading up to 92%. However looking at the importance as measured by the Gini index it is clear that molecular weight, and thus sequence length has played a big part in exceeding accuracy normally achieved in protein classification. However the baseline accuracy of 86% achieved using the amino acid frequencies alone suggest that random forest can classify between the two well. This is not so surprising given that in each of the two sub families of antifreeze proteins, many of the individual proteins will have evolved from a large range of bacterial proteins, each with their own distribution of amino acids. It is an interesting observation that functional groups are mainly absent from the Gini importance as the antifreeze and antifreeze-like proteins share similar functions.

V. Result and Discussion

A. *HMM results*

Testing of proteins against a given HMM is performed by aligning the sequence to the model and scoring it via the Viterbi algorithm. The protein is then scored and expressed a log-odds ratio of the probability of the protein belonging to that HMM divided by the null model. The HMM results are shown in Fig. 5 for transmembrane proteins and Fig. 6 for antifreeze proteins. The results is a bit score, in which for each family PFAM suggests a unique threshold score to be required for a protein to be a possible homology of that protein family.

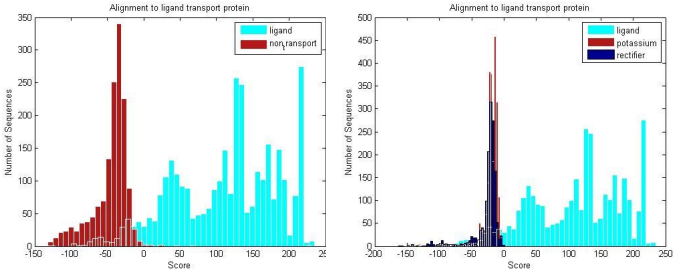

Fig. 5. *Left:* Non-transport and ligand proteins aligned to the ligand HMM model. *Right:* the three different sub-types of transmembrane proteins aligned to the ligand HMM model.

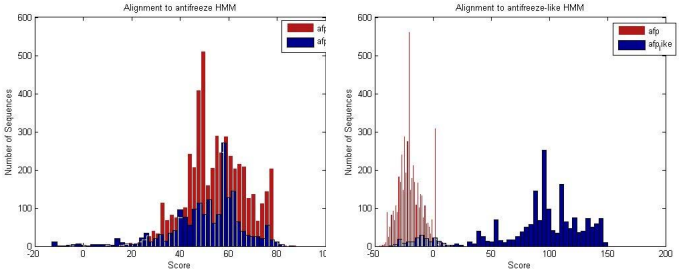

Fig. 6. *Left:* Non-transport and ligand proteins aligned to the ligand HMM model. *Right:* the three different sub-types of antifreeze proteins aligned to the ligand HMM model.

Testing transmembrane proteins against respective HMMs yield high classification accuracy, not only against non-transmembrane proteins, but distinguishing between their sub-types despite similar structure and function. This is not suprising however, as this is because the transmembrane proteins between subfamilies share a lot of sequence similarity, but the HMM models are very sensitive to amino acid changes in certain positions, and vary according to the type of subsitution. On the other hand, because antifreeze proteins evolve from many different ancestors, and generally converge to functional similarity as opposed to sequence similarity, the HMM model struggles to distinguish between antifreeze and antifreeze-like proteins when aligned against the pure antifreeze HMM. The fact that this is not the case when aligned to the antifreeze-like model possibly suggests the antifreeze-like HMM requires any protein to have functional properties that largely reflect the variety of functional groups present within the family, where the antifreeze HMM only requires sub-sequences within the protein responsible for antifreeze functionality - a feature shared between both sub-families. The fact that antifreeze-like proteins scored against the antifreeze-like model yields high variation in the distribution of log-odds scores, yet shows a lower distribution when scored against the antifreeze HMM informs us that a functional, rather then purely sequenced based approach to classifying certain amino acid families may often be needed in families of low sequence homology.

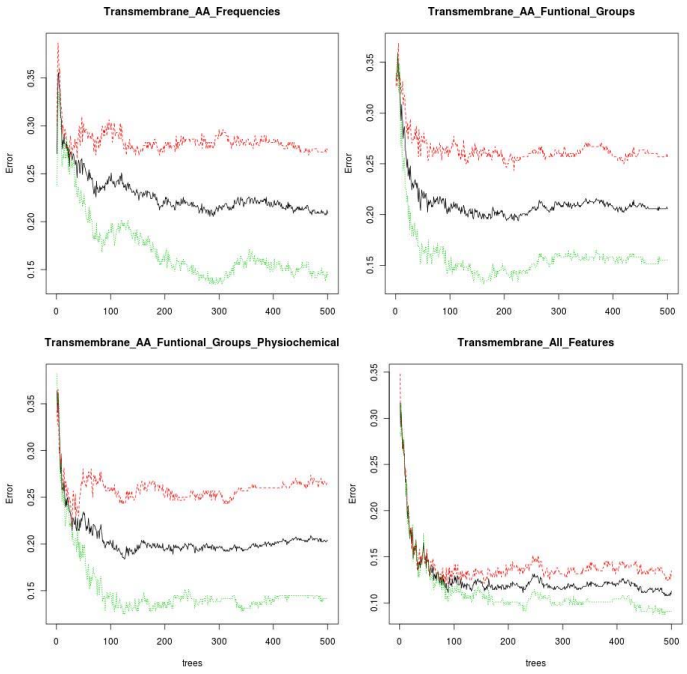

Fig. 7. The OOB error of RF training on transmembrane proteins by using different sets of features. *Top Left*: Amino acids; *Top Right*: Amino acids, and functional groups; *Bottom Left*: Amino acids, functional groups, and chemical properties; *Bottom Right*: Entire feature sets. The black curves are the OOB error which represents the error of proteins used outside of the training process from both classes, the green curves denote flase positives, while red ones are false negatives.

B. Random Forest results

Random Forest however was consistent over both transmembrane and antifreeze classification, however it should be noted that RF is a discriminative process in which classification is performed against a negative training set, and as such a comparison between HMM and RF classification is not explicit here. The results in Table I are listed as the four main feature sets are cumulatively introduced into the training/testing process.

TABLE I: Cumulative classification

Features	Transmembrane	Antifreeze
Amino acids	62.5%	86.9%
Chemical properties	72.5%	89.1%
Functional groups	75%	89.1%
Secondary Structure	85%	91.2%

The Fig. 7 and Fig. 8 show the OOB error of RF training on transmembrane proteins and antifreeze proteins with four different sets of features respectively. The black curves are the OOB error which represents the error of proteins used outside of the training process from both classes. The green curves denote false positive rates, while red ones are false negatives.

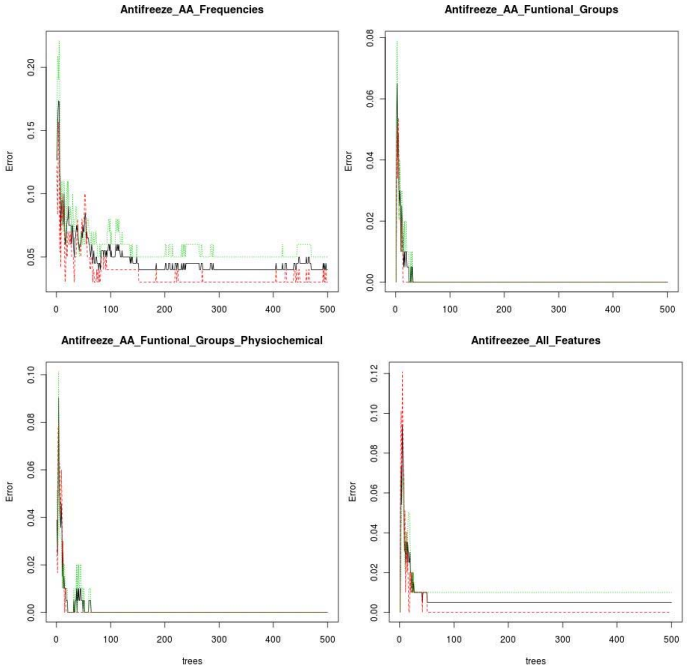

Fig. 8. The OOB error of RF training on antifreeze proteins by using different sets of features. *Top Left*: Amino acids; *Top Right*: Amino acids, and functional groups; *Bottom Left*: Amino acids, functional groups, and chemical properties; *Bottom Right*: Entire feature sets. The black curves are the OOB error which represents the error of proteins used outside of the training process from both classes, the green curves denote flase positives, while red ones are false negatives.

The class error in non-transmembrane and antifreeze-like proteins are particular high compared to that of the class error of classifying transmembrane proteins and thus shows high sensitivity. In particular it can be seen in Fig. 7 that introducing secondary structure into the classification process for transmembrane proteins vastly increases the class error of non-transmembrane proteins. This is expected as a prominent biological features of transmembrane proteins are their rigid structure. The results are not only useful to obtain the accuracy of ranfom forest classification, but it can inform what features are needed to obtain near-optimal results. For example, the classification for antifreeze proteins as a whole shows high accuracy from a small amount of features, even using amino acid frequencies alone. The classification does not show as much improvement form introducing more features, and the classification process is optimal using only 50 trees. The error reduces dramatically as secondary structure is added into the transmembrane classification, as in general transmembrane proteins have rigid secondary structure compared to non-transmembrane proteins.

C. Discussion

The aim of this work was to provide an insight of machine learning in the context of protein classification, in particular the random forest and hidden markov model algorithms. Random forest uses meta data extracted from protein sequences to split the data into user-defined classes, where HMM builds a statistical model from directly aligning protein sequences of known homology, and new sequences are then aligned to the model. There is a large focus in bioinformatics on how various machine learning algorithms compare to each other in terms of classification accuracy, but just as was illustrated in the differences between local and global alignments of sequences (Needleman-Wunsch [] and Smith-Waterman algorithms), each machine learning algorithm should also be viewed on its individual merits and what they can offer.

The fact that profile HMM is built from directly aligning sequences to a model built by aligning *many* sequences provides a classification tool that specialises in finding the local segments of protein sequences that are conserved through evolution and thus provide a way to find distant homologue that have diverged through evolutionary processes. Another aspect of HMM is that a log-odds scoring system provides a metric that is far more expansive than simple and discrete yes/no labels used in Random Forests. This can provide biologists a way of focussing on proteins which do not classify as well and explore through other methods to discriminate them against a family. Random forests on the other hand excel at discriminating between two groups of protiens and are not constrained to classifying on raw sequence data. The ability to extract and train on a vast array of features allows for customisation that can tailor the process based on prior knowledge of the protein groups. The features used to best split the data are computed via the Gini index (or permutation importance is an alternative method) and so provide an insight into what features were used to classify the data. This is an a particularly useful and intutive feature of random forests and can not only be used to verify that important features *expected* to be found within certain families are part of the classifier, but also could potentially provide a platform for investigating features picked out by random forest that were not necessarily thought to be important. One further conclusion from this work is that these two different approaches can perhaps be combined to perform protein classification, e.g. a two-step process where a test protein may score well against a group of HMM can then be used to discriminate which one is the most likely using Random Forest.

REFERENCES

[1] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," *Journal of Molecular Biology*, vol. 157, no. 1, pp. 105 – 132, 1982.

[2] S. F. Altschul and W. Gish, "Local alignment statistics," in *Computer Methods for Macromolecular Sequence Analysis*, ser. Methods in Enzymology, R. F. Doolittle, Ed. Academic Press, 1996, vol. 266, pp. 460 – 480.

[3] W. R. Taylor, "Identification of protein sequence homology by consensus template alignment," *Journal of Molecular Biology*, vol. 188, no. 2, pp. 233 – 258, 1986.

[4] S. Henikoff, "Scores for sequence searches and alignments," *Current Opinion in Structural Biology*, vol. 6, no. 3, pp. 353 – 360, 1996.

[5] S. R. Eddy, "Profile hidden markov models." *Bioinformatics*, vol. 14, no. 9, pp. 755–763, 1998.

[6] E. L. L. Sonnhammer, S. R. Eddy, E. Birney, A. Bateman, and R. Durbin, "Pfam: Multiple sequence alignments and hmm-profiles of protein domains," *Nucleic Acids Research*, vol. 26, no. 1, pp. 320–322, 1998.

[7] K. K. Kandaswamy, K.-C. Chou, T. Martinetz, S. Mller, P. Suganthan, S. Sridharan, and G. Pugalenthi, "Afp-pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties," *Journal of Theoretical Biology*, vol. 270, no. 1, pp. 56 – 62, 2011.

[8] P. L. Davies and B. D. Sykes, "Antifreeze proteins," *Current opinion in structural biology*, vol. 7, no. 6, pp. 828–834, 1997.

[9] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. L. Sonnhammer, J. Tate, and M. Punta, "Pfam: the protein families database," *Nucleic Acids Research*, vol. 42, no. D1, pp. D222–D230, 2014.

[10] M. H. Saier, V. S. Reddy, D. G. Tamang, and . Vstermark, "The transporter classification database," *Nucleic Acids Research*, vol. 42, no. D1, pp. D251–D258, 2014.

[11] L. J. McGuffin, K. Bryson, and D. T. Jones, "The psipred protein structure prediction server," *Bioinformatics*, vol. 16, no. 4, pp. 404–405, 2000.

[12] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa, "Aaindex: amino acid index database, progress report 2008," *Nucleic Acids Research*, vol. 36, no. suppl 1, pp. D202–D205, 2008.

[13] M. Kanehisa and S. Goto, "Kegg: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

[14] S. Unterreitmeier, A. Fuchs, T. Schffler, R. G. Heym, D. Frishman, and D. Langosch, "Phenylalanine promotes interaction of transmembrane domains via gxxxg motifs," *Journal of Molecular Biology*, vol. 374, no. 3, pp. 705 – 718, 2007.

[15] J. U. Bowie, "Helix packing in membrane proteins," *Journal of Molecular Biology*, vol. 272, no. 5, pp. 780 – 789, 1997.

[16] N. Li, C. A. Andorfer, and J. G. Duman, "Enhancement of insect antifreeze protein activity by solutes of low molecular mass." *The Journal of Experimental Biology*, vol. 201, no. 15, pp. 2243–51, 1998.