Bioinformatics, 31(21), 2015, 3429–3436
doi: 10.1093/bioinformatics/btv345
Advance Access Publication Date: 30 June 2015
Original Paper

OXFORD

Sequence analysis

ProFET: Feature engineering captures high-level protein functions

Dan Ofer and Michal Linial*

Department of Biological Chemistry, Institute of Life Sciences, The Edmond J. Safra Campus, The Hebrew University of Jerusalem, Givat Ram, 91904, Israel

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on March 4, 2015; revised on May 3, 2015; accepted on May 29, 2015

Abstract

Motivation: The amount of sequenced genomes and proteins is growing at an unprecedented pace. Unfortunately, manual curation and functional knowledge lag behind. Homologous inference often fails at labeling proteins with diverse functions and broad classes. Thus, identifying high-level protein functionality remains challenging. We hypothesize that a universal feature engineering approach can yield classification of high-level functions and unified properties when combined with machine learning approaches, without requiring external databases or alignment.

Results: In this study, we present a novel bioinformatics toolkit called ProFET (Protein Feature Engineering Toolkit). ProFET extracts hundreds of features covering the elementary biophysical and sequence derived attributes. Most features capture statistically informative patterns. In addition, different representations of sequences and the amino acids alphabet provide a compact, compressed set of features. The results from ProFET were incorporated in data analysis pipelines, implemented in python and adapted for multi-genome scale analysis. ProFET was applied on 17 established and novel protein benchmark datasets involving classification for a variety of binary and multi-class tasks. The results show state of the art performance. The extracted features' show excellent biological interpretability. The success of ProFET applies to a wide range of high-level functions such as subcellular localization, structural classes and proteins with unique functional properties (e.g. neuropeptide precursors, thermophilic and nucleic acid binding). ProFET allows easy, universal discovery of new target proteins, as well as understanding the features underlying different high-level protein functions.

Availability and implementation: ProFET source code and the datasets used are freely available at https://github.com/ddofer/ProFET.

Contact: michall@cc.huji.ac.il

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The most used approaches in protein classification rely on distance measures between sequences based on various alignment methods (e.g. Smith-Waterman, BLAST). With the growth in the amounts and diversity of protein sequences, more sophisticated methods have been introduced (e.g. PSSM, Profile-Profile, HMM-HMM) (Jaakkola *et al.*, 2000; Soding, 2005). These methods are based on multiple sequence alignments for improving remote homologs detection (Edgar and Sjolander, 2004; Karplus *et al.*, 1998). Incorporating 3D-structure as a seed for the statistical models further improved the quality of protein domains and families (e.g. Pfam) (Finn *et al.*, 2014; Sonnhammer *et al.*, 1997). Currently, there are ~27 000 such models (InterPro, Mulder and Apweiler, 2007) that cover 83% of all sequences in UniProtKB (2014_10). Function assignment is gained from mapping InterPro models to Gene Ontologies (i.e. InterPro2GO). An alternative model-free approach

was proposed (Portugaly *et al.*, 2002). The assessment of large-scale automatic protein functional annotations (Radivojac *et al.*, 2013; Rost *et al.*, 2003) and the contribution of alternative approaches toward this task have been extensively discussed (e.g. Valencia, 2005).

Despite the strength of the model-based methods, in many instances the local sequence-based methods fail to reliably assign a function (Rentzsch and Orengo, 2009). This is best demonstrated by the limitation in classifying proteins by their 3D-folds (Greene *et al.*, 2007; Todd *et al.*, 2005). Notably, the classification of some biological niches is especially suited for feature representation. For example, routine annotation tools fail to confidently assign function for bioactive peptides and short proteins (Naamati *et al.*, 2009). A number of previous studies focus on feature extraction from whole protein sequences (Cao *et al.*, 2013; Ding and Dubchak, 2001; Dubchak *et al.*, 1995; Nanni *et al.*, 2014; van den Berg *et al.*, 2014) as a starting input for machine learning (ML) approaches. Structural benchmark from SCOP and CATH (Lewis *et al.*, 2013) are frequently used to assess the predictive ML methods. Specialized predictors have been presented for structural tasks including secondary structure, solvent accessibility, stability, disordered regions, domains and more (Cai *et al.*, 2001; Cheng *et al.*, 2005; Ding and Dubchak, 2001). ML approaches have proven suitable to classify protein properties beyond their 3D-structure. SVMProt was tested on preselected 50 functional families from Pfam (Chou and Cai, 2003). Naive biophysical features classification outperformed simple sequence-based methods for a number of protein families (Varshavsky *et al.*, 2007). However, the most likely advantage of the feature and pattern-based ML approach is toward high-level functionality (e.g. Pe'er *et al.*, 2004). Examples for such predictions include protein–protein interactions (Bock and Gough, 2001; Cheng and Baldi, 2007), discriminating outer membrane proteins (Gromiha and Suwa, 2005), membrane topology (Nugent and Jones, 2009), subcellular localization (Hua and Sun, 2001) and more. The strongest features learned by the ML classifiers often expose biologically important motifs (Leslie *et al.*, 2004).

In this study, we focus on the ability of elementary biophysical features together with a rich set of engineered representation of proteins to classify high-level protein functions. These features are suited for both supervised and unsupervised classification. Our goal is to illustrate the importance of ProFET (Protein Feature Engineering Toolkit) as a 'one size fits all' framework for representing whole protein sequence. We present a universal, modular workflow for protein function classification: (i) feature generation and extraction from primary sequences (ProFET). (ii) Application of the extracted features in a ML framework for binary or multi-class partition. (iii) Presentation of discriminative classification power. (iv) Identification of patterns and features that underlie the successful classification ('Feature Selection').

2 Methods

2.1 Protein databases and datasets

In gathering the protein sets in this study, we used datasets made available by (i) custom sets gathered from public databases such as UniProtKB (Wu *et al.*, 2006) and SCOPe (Fox *et al.*, 2014) and (ii) benchmarks extracted from publications. For both resources, we applied CD-Hit and USearch (Edgar, 2010) to remove redundant sequences according to a predefined % of sequence identity. As a rule, we used only classes that contain a minimal number of samples per group (typically 40, after redundancy removal). Sequences with unknown amino acid (AA), errors or sequences that are shorter than 30 AA were removed. We included in the analysis the most recent SCOP classification (2.05, 71015 PDB entries pre-filtering) as some literature-based benchmarks from SCOP were outdated (Chandonia *et al.*, 2004).

2.1.1 Specialized protein functions

- Neuropeptide precursors (NPPs): The keyword 'neuropeptide' is acquired from SwissProt (SWP) and UniRef90 representatives. We removed proteins that contain the terms 'fragment' and 'receptor'.
- Ribosomal proteins: Acquired from SWP and partitioned to Archea, Bacteria and Eukarya. Redundancy filter was set to 20–40% identity (according to the set size).
- Thermophilic proteins: The ThermoPred benchmark dataset (Lin and Chen, 2011) was used, with a further redundancy removal (at 40% identity threshold).

2.1.2 Cellular localizations

- LocTree3 benchmark (Yachdav *et al.*, 2014) for Eukarya and Bacteria were used. Filtered at 40% identity within each class.
- Mammalian subcellular localization: Protein-organelle pairs are acquired from SWP.
- Uncultured bacterium. Sequences extracted from UniProtKB and mapped to keyword annotations for major cellular compartments (membrane, cytoplasm, ribosome). Filtered at 50% identity according to UniRef clusters.

2.1.3 Structural-based classifications

- SCOPe (Release 2.05, February 2015) (Fox *et al.*, 2014). Classes and folds were defined by SCOP, with 25% or 10% sequence identity filter (8514 and 6721 sequences, respectively).
- SCOPe (Release 2.05, February 2015) 'selected class' defined by the SCOP class (marked a–k), with classes c,d removed. We also apply as a benchmark classes 'a,b,f,g' at 25% sequence identity filter. Classes a,b,c,f,g were tested following redundancy removal at extremely low identity level (10%). The classes that were not included had small number of folds in each.

2.1.4 Nucleic acids binding proteins

- DNA-binding proteins. Benchmark dataset from DNA binder (Kumar *et al.*, 2009).
- RNA-binding proteins. Benchmark dataset from BindN (Wang *et al.*, 2010).

2.1.5 Viral properties and classes

- Virus-host pairs: Acquired from SWP. The set include all viral proteins partitioned by the kingdom of the hosts. Redundancy filtration (at 40% identity) was performed on the viral proteins but not on the hosts.
- Capsids: Compilation of two sets of all viral capsid proteins annotated by SWP: (i) Classes according to host type. (ii) Classes according to viral replication mode.

The datasets and sequences used are all freely provided online: https://github.com/ddofer/ProFET.

2.2 Features

All features extracted by ProFET are directly derived from the protein sequence and do not require external input (Saeys *et al.*, 2007). The software packages required for ProFET are part of the scientific Python distribution. Properties relying on external predictors (e.g. the 3D structural fold, secondary structure) are not included by default. However, users can trivially add additional features via the 'FeatureGen' script. ProFET can also generate a pre-defined set of default features for consistency in evaluation and ease of use, callable from the command-line.

The features that are described below can be restricted to a segment of a protein (e.g. each individual third of a sequence). We support two versions for a subsequence analysis: (i) relative portions and (ii) fixed lengths. The activation of global feature extraction combined with segmental consideration is advantageous. It is motivated by the atypical composition of different segments of numerous protein classes, e.g. the signal peptides, flexible N-terminal linker regions, C-terminal portions of membranous kinases and GPCR receptors, disordered regions and more.

The categories of features currently implemented in ProFET are as follows.

2.2.1 Biophysical quantitative properties

i. Molecular weight (in Da)
ii. Sequence length (in AA)
iii. pH(I), the isoelectric point
iv. Net Charge at various pH(I)s.
v. Aromaticity the relative frequency of Phe, Trp, Tyr.
vi. Instability index, an estimate for the stability of a protein *in vitro* (Gasteiger *et al.*, 2003).
vii. GRAVY (Grand Average of Hydropathy), the sum of hydropathy values of all AA, divided by the number of AA in the analyzed sequence (Kyte and Doolittle, 1982).
viii. Aliphatic index, the relative volume occupied by aliphatic side chains (Ala, Val, Ile and Leu) (Gasteiger *et al.*, 2003).

Most of these properties were based on the Expasy proteomics collection (Gasteiger *et al.*, 2003). The important of these elementary global features has been previously validated (Varshavsky *et al.*, 2007).

2.2.2 Letter-based features

i. AA composition (single or di-peptide)
ii. Overlapping K-mers.
iii. 'Mirror' K-mers. It accounts for K-mers of various combinations of 'grouped' AA. For example, lysine-arginine appearance (KR) is grouped together with RK.
iv. Reduced AA alphabets. Grouping of AA secures a compact representation. We include a large number of such alphabets from various sources (Murphy *et al.*, 2000; Peterson *et al.*, 2009) and some novel alphabet representations of size 14 and 8 (Ofer_14 and Ofer_8, respectively). For the 14 AA representation, the grouping is for KR, TS and LIVM. For the 8 AA representation, the grouping is for FYW, ALIVM, RKH, DE and STNQ. The other AA remain in the uncompressed representation.

2.2.3 Local potential features

i. Potential post-translational modification (PTM) sites. We included motifs implemented as regular expressions, including those for 'known short motif' dibasic cleavage model (X-X-Lys-[Lys or Arg], X-X-Arg-Arg, Arg-X-X-[Lys or Arg]; where X denotes any AA (Southey *et al.*, 2006; Veenstra, 2000). Others include N-glycosylation and Asp or Asn hydroxylation sites. We included Cysteine spacer motif that captures the tendency of Cys to appear in a minimal window (Naamati *et al.*, 2009). Additional PTM motifs collected from ELM (Dinkel *et al.*, 2012) were not implemented.
ii. Potential Disorder (FoldIndex). Local regions of disorder are predicted using the naive FoldIndex (Prilusky *et al.*, 2005) and TDP-IDP methods (Campen *et al.*, 2008; Klus *et al.*, 2014). FoldIndex predicts the disorder as a function of the hydrophobic potential and net charge.

2.2.4 Information-based statistics

These features aim to capture the non-randomly distribution of each AA in the sequence, based on the concept of information entropy.

The information-based features used are:

i. Total entropy per letter, as a whole
ii. The binary autocorrelation
iii. Autocorrelation with Selected letters. For example, K, R or C is denoted as '1' and the rest as '0'. Lag is then computed. For details, see Ofer and Linial (2014).

2.2.5 AA scale-based features

AA propensity scales map each AA to a quantitative value that represents physicochemical or biochemical properties, such as hydropathicity or size. These scales can then be used to represent the protein sequence as a time series, typically using sliding windows of different sizes and to extract additional features.

ProFET includes a wide array of scales, ranging from the established propensities for hydrophobicity and flexibility/B-factors (acquired from Expasy), to 'optimal' and maximally independent derived scales (Atchley *et al.*, 2005; Georgiev, 2009).

Features derived from these scales include:

i. Averages for the sequence as a whole, for different window sizes.
ii. Quartile averages (e.g. top 25%).
iii. Maximum and minimum values for a given scale and window-size along the entire sequence.
iv. Autocorrelation.

2.2.6 Transformed CTD features (Dubchak *et al.*, 1995)

We implemented the Dubchak and ProFEAT CTD features (hydrophobicity, normalized Van der Waals volume, polarity, polarizability, charge, secondary structure and solvent accessibility hydrophobicity, normalized Van der Waals volume, polarity, polarizability, charge, secondary structure and solvent accessibility) (Dubchak *et al.*, 1995; Li *et al.*, 2006). Code from Spice (van den Berg *et al.*, 2014) and (Cao *et al.*, 2013) was also integrated. An additional subdivision of disorder propensity was adapted from Composition Profiler (Vacic *et al.*, 2007): 1:'ARSQEGKP', 2:'ILNCFYVW' and 3:'DHMT'.

The features used are:

i. Composition (C) is the number of AA of a particular property divided by the total number of AA.
ii. Transition (T) is the number of transitions from a particular property to different property, divided by (total number of AA − 1).
iii. Distribution (D) captures is the chain length within which the first 25%, 50%, 75% and 100% AA of a particular property are located.

2.3 Evaluation

The power of any of the predictor proposed is tested by several routinely used evaluation methods. We measure the performance for the binary and multiclass tasks with the same metrics: F1 score (the weighted average of the precision and recall) and Accuracy (Acc).

These parameters are defined as:

- $F1 = 2*TP/(2TP + FP + FN)$
- $Ac = (TP + TN)/(TP + TN + FP + FN)$

TP represents the number of the correctly recognized proteins. FP, the number of proteins wrongly identified and FN the number of proteins missed. Performance is evaluated using cross-validation.

Specifically, multiple rounds of randomized stratified cross validation ('Stratified Shuffle Split'), with 18% holdout for each iteration (unless mentioned otherwise). Features were filtered prior to cross validation and testing using a simple univariate filter for statistical significance ($a \leq 0.01$, Bonferroni multiple testing family wise error rate corrected; analysis of variance one-way *F*-test). This prefiltering step at the cross validation phase had a negligible impact on the overall performance (not shown).

2.4 Feature selection

A wide array of methods for supervised and unsupervised feature selection can be applied to identify the best features, implemented with the superlative Scikit learn toolkit (Abraham *et al.*, 2014). These include wrapper methods—Random Feature Elimination (Ozcift, 2012), model-based filtering [e.g. support vector machine (SVM) classifiers with a L1 Loss penalty, for sparse coefficients], statistical filtering, stability selection, PCA, etc.

In the test cases, we used the RFE method, combined with an underlying non-linear ensemble of classifiers (Random forests). The underlying principle is iterative fitting of the classifier on the data, with the weakest features being pruned at each of the iterations (Abraham *et al.*, 2014). We examined the selected features, and the model classification performance with the reduced set of features, and show novel, interpretable features, as well as excellent retained performance.

3 Results

3.1 ProFET outline

We introduce two test cases to illustrate the potential of ProFET to provide a generic platform for analyzing the basis of high-level functionality in proteins.

Classifying thermophile proteins was used as a test case for a binary classification of functionality that is not explicitly derived from the sequence. Classifying neuropeptide (NP) hormone precursors serves to assess the classification of poorly studied protein niche (Karsenty *et al.*, 2014). We generalize the approach to a range of from subcellular localization to viral phylogeny tasks (see Section 2.1.1–2.1.5). In all the illustrated cases, ProFET was used as a generic framework for feature extraction and prediction. External information that is often available (e.g. the family PSSM, GO annotation, structural prediction and disorder predictors) was not included.

The workflow is composed of modular sections (Fig. 1)

1. ProFET: Feature extraction from any protein sequences. Extracted features can be analyzed independently (suitable for ML analysis or unsupervised tasks) or discriminatively (i.e. seeking contrast between groups of proteins).
2. Model Selection: The features are used to train and tune different ML models. For any given performance metric (e.g. precision), the optimal model and hyper-parameters are selected.
3. Performance Report: Classification performance is measured for a given model and dataset, using cross-validation.
4. Feature Selection: Informative features are selected and their importance measured using different methods. These methods include the statistical significance, wrapper methods, model-based selection, stability selection and more.
5. New sequences can be predicted using a trained ML model. This can be applied via the feature extraction pipeline or with a selected smaller subset of the selected features.

3.2 ProFET workflow—case studies

We selected three datasets to illustrate the performance of ProFET and its workflow (Fig. 1).

3.2.1 Positive–negative protein sets

Set 1: Thermophiles are proteins that function under high temperature. Given the extreme environmental conditions, we expect to detect biophysical signatures in these proteins underlying their thermostability. We used a benchmark dataset of 915 thermophilic and 793 non-thermophilic (Mesophile) proteins that were further filtered to insure $< 40\%$ sequence identity between sequences within each group (Lin *et al.*, 2005).

Set 2: NPPs are pre-pro-polypeptide precursors of NPs. These are secreted proteins. Routine sequence alignment-based methods are insufficient to identify the immensely diverse NPs. In compiling a dataset, we used as a negative set a collection of proteins with Signal peptides, which lacked validated TMD (and therefore, most likely to be secreted). We keep the same (atypical) range of lengths to match the labeled NPPs. Both the positive and negative datasets have Signal peptides confirmed and cleaved using SignalP (Petersen *et al.*, 2011). The negative (non-NPP) dataset was filtered using Usearch (Edgar, 2010), so that proteins in the negative set shared no sequence similarity (cutoff of 10% identity was applied). The final dataset held 2309 negatives and 1269 NPPs. Note that in the case of NPPs, we expect many unidentified NPP peptides among the proteins in the negative set.

Set 3: Uncultured bacteria account for ~250 K proteins in UniProtKB (Wu *et al.*, 2006). We restricted the test to those having GO annotations for 'ribosome', 'membrane' or 'cytoplasm'. Proteins were filtered for redundancy according to UniRef50 classification,

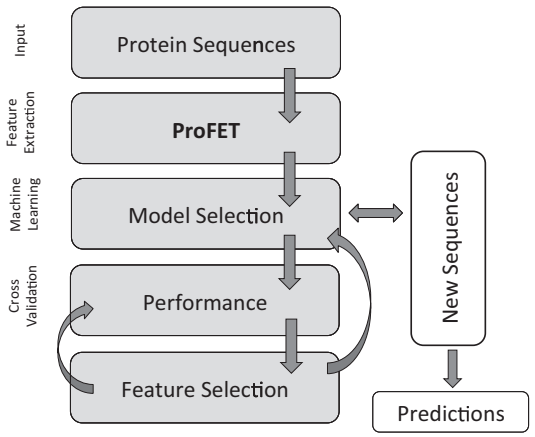

Fig. 1. The ProFET framework: merging machine-learning protocols, cross-validated tuning, feature selection and prediction

leaving 15 995 sequences, 59.2% of them being 'membrane' proteins.

3.2.2 Classification results

For all three sets (as in Section 3.2.1), we obtained almost perfect classification. Classification was performed using a random forest classifier, implemented in Scikit learn (see Section 2). Figure 2A shows the results of the classifications for the set of the Thermophilic proteins and the NPPs as confusion matrices. Results were derived from 10-fold stratified cross validation. In both sets, the number of missed classified (FN and FP) is below 5% for the NPPs (63 and 110 proteins as FN and FP, respectively). For the set of the thermophiles, the missed classifications of the FN and FP reach 6% and 10%, respectively. Figure 2B shows the performance as receiver operating characteristics curves. Performance was measured using an automatically tuned SVM with a radial basis function (RBF) kernel, with 15-fold stratified cross validation. The performance was very high with a FP rate of 0.1 and the AUC for both tests reaching 0.97 (out of a maximum of 1.0).

Uncultured bacteria comprise a set of poorly characterized proteins (Set 3). We trained proteins that mapped three main compartments in bacterium (membrane, cytoplasm, ribosome, total of 15 995 sequences). The localization performance for the multi-class task is very convincing (tested via 12 rounds of stratified shuffle split cross-validation). The F1 score is 0.917 (±0.01 SD); accuracy is 0.916 (±0.01 SD).

We further used a combination of ML approach with naive PSI-Blast search. We activated PSI-Blast (three iterations, default parameters) on sets 1 and set 2 (Thermophiles and NPPs sets). The most

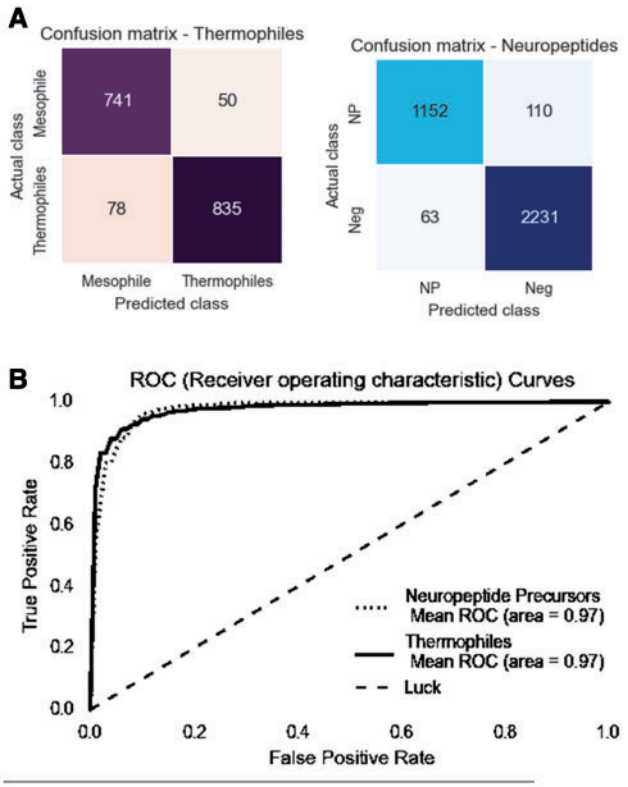

Fig. 2. Performance results for the two datasets used. (A) Confusion matrix of the classifier performance. Results were derived from 10-fold stratified cross-validation. The number of FP and FP is shown for thermophiles (left) and NPs (right). (**B**) AUC (area under receiver operating characteristics curve)

significant *E*-value was used for each sequence as an approximate distance matrix. We then trained a K-nearest-neighbors classifier and recorded the performance. We also used an unsupervised, clustering approach (spectral clustering and K-means) and compared these clusters to the 'true' labels.

Clustering performance was significantly lower than reported (Fig. 2). The best results for the Psi-Blast test were obtained from Spectral clustering model. For the NPP set (total of 3370 proteins), the F1 score is 0.56. The similar analysis for the Thermophile/Mesophile proteins reached F1 score of 0.29 (total of 1708 proteins). To make sure that the poor performance is not dependent on the choice of the ML methodology, we repeated the analysis for a classification by K-nearest neighbors classifier ($k = 1$ or 2). The data were split 80/20 into evaluation and hold-out sets, and the best parameters on the evaluation set were determined by 4-fold cross validation. For the NPP and Thermophile sets, the accuracy on the 'evaluation set' was 62.8% (±0.16 SD) and 48.9% (±0.03 SD), respectively. The F1 score for the hold-out sets were 0.61 and 0.44 for sets 2 and 1, respectively.

3.3 Post-training feature selection

In addition to the success of the predictors, interpretability of the features that best contributed to the performance is a crucial knowledge. Several methods for feature selection can be applied to identify a minimal set of such features. We applied a combination of Random Forests (an ensemble of decision tree classifiers) with the Random Feature Elimination wrapper method.

In each of the iterations, the weakest features are removed and the model is then retrained with the remaining features, until the preselected desired amount of features remains. Performance of the reduced feature set is measured using new splits of the training data and cross validation. Recall that the initial set of (default) generated features included 771 features. The *F*-test filter reduced the number of features to 453 and 544 features for the Thermophiles and NPP sets, respectively.

3.3.1. Thermophilic proteins—informative features

We note the importance of AA composition, particularly of charged and polar AA groups. Of further importance are features involving glutamic acid (E) and glutamine (Q), and the organizational entropy of E and Q. The relevance of these AA was reported (Lin and Chen, 2011; Zhang and Fang, 2007). We note that merely using the AA composition would not have captured many of these features.

The classification performance (F1 score) with just 15 features reached 99.53% of that obtained using all statistically significant features (F1 score = 0.906; 453 features).

3.3.2 NPPs feature—informative features

As opposed to the features dominating the test case of thermophilic proteins, in the case of the NPPs, a smaller set of features dominates, mainly relating to the normalized frequency of putative NPP cleavage sites, according to the 'known motif' model. Further properties of the basic residues Lys (K) and Arg (R) repeat themselves by virtue of entropy, binary autocorrelation (6/15 features) and more. Additional features include protein size (Mw and length) and to a lesser extent some 'structural' properties, such as flexibility ('Flex_min'), and secondary structural propensities—reflecting the importance of availability of the putative cleavage sites and atypical composition of the putative peptides.

Classification performance: (F1 score of the positive class) with just 15 features was 95.85% of that obtained using all statistically significant features [0.945 (∓ 0.01); 544 features].

Figure 3 shows the types of the 15 strongest features for the two test cases. Selected features are ranked by relative importance to the classifier. Feature titles are self-explanatory. For example, 'ofer14KC' specifies the reduced AA alphabet ofer14 (see Section 2.2.2) for grouping of KC in the reduced representation.

3.4 Benchmarks' performance

The workflow applied to our test cases (Section 3.2) was systematically applied to all the datasets. Each set was measured using 15-fold randomized stratified cross-validation. For each iteration, a fraction of the data (18%) is randomly set apart. The framework's automatically selected the performance of the classifier. The term 'Dummy - by majority' applies to a classifier that always picks the majority class. Altogether, we present 15 additional datasets (in addition to the NPPs and Thermophilic proteins). For 76% of the datasets, the accuracy and F1 Scores are above 80%, while for 35%, the accuracy is $>90\%$ (Fig. 4).

The classification performance for DNA and RNA binding proteins meets the state of the art results obtained by special purpose predictors (Wang, etal., 2010). This specialized predictor for DNA and RNA binding proteins relies on the specific evolutionary information (e.g., PSSMs) combined with Support Vector Machine (SVM) (Wang, etal., 2010). 72.42% Accuracy is reported for DNA binding proteins using a random forest model and extensive feature selection (Kumar *et al.*, 2009).

We used the same benchmark data to directly assess the performance. We show (Fig. 4) that our platform reaches a classification success of 0.72 and 0.79 for DNA and RNA binding proteins, respectively. We conclude that excellent performance is achieved by using the default setting of the ProFET workflow.

Five of the benchmarks (Supplementary Table S1, Fig. 4) concern structural SCOP datasets, at the class or fold level. The classification success varies according to the tasks. For example, the success for the SCOP 'selected class' is very high (0.82–0.9), whereas the performance for the fold classification is much lower (0.62–0.65). Note that SCOP 25% and SCOP 10% tasks use the same dataset (SCOPe version 2.05). These sets differ only by the degree of redundancy removal. We found similar levels of accuracy for both sets. The performance (accuracy, F1 score) for all 17 analyzed datasets with respect to the Dummy-majority classifier is shown (Supplementary Table S1).

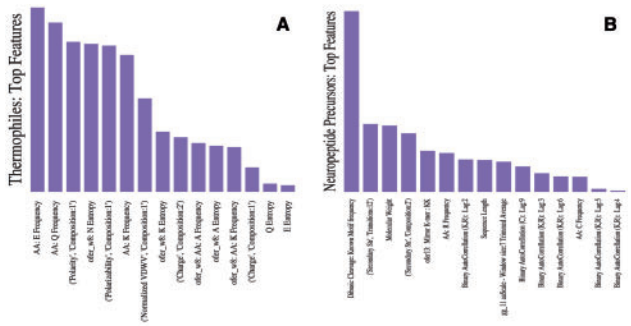

Fig. 3. Top 15 informative features that dominate the successful classification of thermophilic proteins and NPPs

4 Discussion

The main drawbacks in existing sequence-based methods are (i) some functions cannot be detected by sequence-based methods; (ii) current statistical models mostly capture local patterns rather than high-level function and (iii) rare sequences or those that have very few homologs cannot be successfully used for inference or construction of good statistical model.

4.1 Compact representations

In this study, we introduce ProFET as a feature extraction platform that can serve many classification tasks. ProFET was compiled as a flexible tool for any size of protein sequence. Our platform adds to previous studies that use quantitative feature representations for

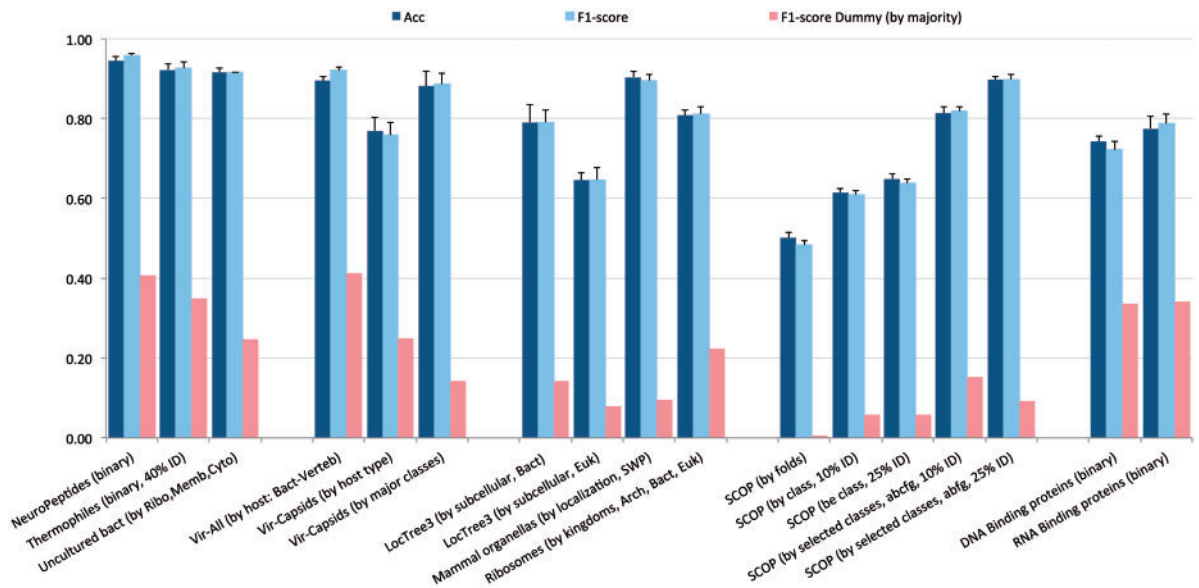

Fig. 4. Classification performance by the Accuracy and F1 score for 17 datasets using the ProFET combined with the ML scheme. Results and SD for the Accuracy (Acc, dark blue) and the F1-score (light blue, middle) are shown. Dummy predictor is a default classifier for the largest class in the dataset (rightmost, coloured pink)

sequences. The communality in these methods is the transformation step in which the protein sequences are converted to hundreds or thousands of features, many of them elementary biochemical and biophysical properties, while others are statistically derived (e.g. frequency of AA and dipeptides).

ProFET includes many novel additions for the elementary representation. For example, features that are based on a reduced alphabets, entropy, high performance AA scales, binary autocorrelation, sequence segmentation, mirror k-mers and more. Many of these features not only improved performance while allowing a compact representation but also expose statistical importance properties in proteins (Fig. 3). The advantage of using reduced alphabet has been noted for 3D-structure representation (Bacardit *et al.*, 2009) and more (Weathers *et al.*, 2004).

ProFET results were the input for ML approaches allowing a rigorous assessment of performance and reaches state of the art results. Recovering the classification success by a small set of top features argues for the power of a compact representation for understanding the features that dominate any specific tasks.

4.2 The user perspective

Several conclusions can be drawn from the results of the classification tasks (Fig. 4):

A. Protein centric analysis: Feature engineering methods presented in this study should be considered a baseline approach for whole protein rather than protein domains. Most of our knowledge from 3D structure and evolution relies on the properties of domains within proteins. We propose the feature engineering as a complementary approach to the domain-centric one.

B. 'One size fits all': Features that are included in ProFET are highly relevant to a broad range of proteins. This is in contrast to methods that customize features for a specific task. The ProFET pipeline provides a default set of features that is suitable for many classification tasks. Therefore, ProFET eliminate the need to duplicate the effort for feature extraction.

C. Flexibility of use: Our presented pipeline accepts a single sequence, combined files, multiple files or a directory. It automatically labels the input into classes (if desired) and normalizes the features (if desired). Thus, any user can use ProFET to set the desired combination of features, representations and normalization. From the point of view of the user, several considerations were taken:

- Our pipeline handles FASTA files and stores them as labeled CSVs.
- We use state of art, open source, freely available python data science tools (such as Pandas, scikit-learn, biopython) (Cock *et al.*, 2009).
- Easy to add new features using a standardized format.
- Our framework includes details on the features as part of the data pipeline so results are interpretable.
- Our code is available for academic and non-commercial use, under the GNU 3 license.

We provide a large collated resource for feature extraction. Thanks to the modular design of ProFET, adding and tinkering with features is trivial. Users of ProFET can decide to focus, remove or expand any subset of the features (e.g. k-mer lengths). ProFET allows tuning of any number of parameters in the feature generation pipeline, e.g. the AA scales to use and the elementary window size for extracting properties. In addition, features can be extracted locally from the N' terminals or the C'-terminals or from an arbitrary segment of the protein.

In summary, the approach presented here is suitable and powerful for application towards modern approach for ML especially in the emerging field of Deep Learning and unsupervised learning of feature representations. These features can easily experimented with allowing additional applications of biological insight to the task of feature engineering.

Acknowledgements

We thank Michael Doron for extensive collaboration, aid and programming expertise in setting up the framework. Nadav Rappoprt supported Psi-Blast comparisons. We thank Nadav Rappoport, Nadav Brandes and Kerem Wainer for fruitful discussions. The project is part of the ELIXIR infrastructure.

Conflict of Interest: none declared.

References

Abraham,A. *et al.* (2014) Machine learning for neuroimaging with scikit-learn. *Front. Neuroinform.*, **8**, 14.

Atchley,W.R. *et al.* (2005) Solving the protein sequence metric problem. *Proc. Natl. Acad. Sci. USA*, **102**, 6395–6400.

Bacardit,J. *et al.* (2009) Automated alphabet reduction for protein datasets. *BMC Bioinformatics*, **10**, 6.

Bock,J.R. and Gough,D.A. (2001) Predicting protein–protein interactions from primary structure. *Bioinformatics*, **17**, 455–460.

Cai,Y.D. *et al.* (2001) Support vector machines for predicting protein structural class. *BMC Bioinformatics*, **2**, 3.

Campen,A. *et al.* (2008) TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. *Protein Pept. Lett.*, **15**, 956–963.

Cao,D.S. *et al.* (2013) propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics*, **29**, 960–962.

Chandonia,J.M. *et al.* (2004) The ASTRAL Compendium in 2004. *Nucleic Acids Res.*, **32**, D189–D192.

Cheng,J. and Baldi,P. (2007) Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics*, **8**, 113.

Cheng,J. *et al.* (2005) SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res.*, **33**, W72–W76.

Chou,K.C. and Cai,Y.D. (2003) Prediction and classification of protein subcellular location-sequence-order effect and pseudo amino acid composition. *J. Cell. Biochem.*, **90**, 1250–1260.

Cock,P.J. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.

Ding,C.H. and Dubchak,I. (2001) Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, **17**, 349–358.

Dinkel,H. *et al.* (2012) ELM—the database of eukaryotic linear motifs. *Nucleic Acids Res.*, **40**, D242–D251.

Dubchak,I. *et al.* (1995) Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl. Acad. Sci. USA*, **92**, 8700–8704.

Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Edgar,R.C. and Sjolander,K. (2004) COACH: profile-profile alignment of protein families using hidden Markov models. *Bioinformatics*, **20**, 1309–1318.

Finn,R.D. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.

Fox,N.K. *et al.* (2014) SCOPe: structural classification of proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.*, **42**, D304–D309.

Gasteiger,E. *et al.* (2003) ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.*, **31**, 3784–3788.

Georgiev,A.G. (2009) Interpretable numerical descriptors of amino acid space. *J. Comput. Biol.*, **16**, 703–723.

Greene,L.H. *et al.* (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res.*, **35**, D291–D297.

Gromiha,M.M. and Suwa,M. (2005) A simple statistical method for discriminating outer membrane proteins with better accuracy. *Bioinformatics*, **21**, 961–968.

Hua,S. and Sun,Z. (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17**, 721–728.

Jaakkola,T. *et al.* (2000) A discriminative framework for detecting remote protein homologies. *J. Comput. Biol.*, **7**, 95–114.

Karplus, K. *et al.* (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.

Karsenty,S. *et al.* (2014) NeuroPID: a classifier of neuropeptide precursors. *Nucleic Acids Res.*, **42**, W182–W186.

Klus,P. *et al.* (2014) The cleverSuite approach for protein characterization: predictions of structural properties, solubility, chaperone requirements and RNA-binding abilities. *Bioinformatics*, **30**, 1601–1608.

Kumar,K.K. *et al.* (2009) DNA-Prot: identification of DNA binding proteins from protein sequence information using random forest. *J. Biomol. Struct. Dyn.*, **26**, 679–686.

Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.

Leslie,C.S. *et al.* (2004) Mismatch string kernels for discriminative protein classification. *Bioinformatics*, **20**, 467–476.

Lewis,T.E. *et al.* (2013) Genome3D: a UK collaborative project to annotate genomic sequences with predicted 3D structures based on SCOP and CATH domains. *Nucleic Acids Res.*, **41**, D499–D507.

Lin,H. and Chen,W. (2011) Prediction of thermophilic proteins using feature selection technique. *J. Microbiol. Methods*, **84**, 67–70.

Lin,C. *et al.* (2013) Hierarchical classification of protein folds using a novel ensemble classifier. *PloS One*, **8**, e56499.

Lin,K. *et al.* (2005) A simple and fast secondary structure prediction method using hidden neural networks. *Bioinformatics*, **21**, 152–159.

Mulder,N. and Apweiler,R. (2007) InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods Mol. Biol.*, **396**, 59–70.

Murphy,L.R. *et al.* (2000) Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Eng.*, **13**, 149–152.

Naamati,G. *et al.* (2009) ClanTox: a classifier of short animal toxins. *Nucleic Acids Res.*, **37**, W363–W368.

Nanni,L. *et al.* (2014) An empirical study of different approaches for protein classification. *ScientificWorldJournal*, **2014**, 236717.

Nugent,T. and Jones,D.T. (2009) Transmembrane protein topology prediction using support vector machines. *BMC Bioinformatics*, **10**, 159.

Ofer,D. and Linial,M. (2014) NeuroPID: a predictor for identifying neuropeptide precursors from metazoan proteomes. *Bioinformatics*, **30**, 931–940.

Ozcift,A. (2012) Enhanced cancer recognition system based on random forests feature elimination algorithm. *J. Med. Syst.*, **36**, 2577–2585.

Pe'er,I. *et al.* (2004) Proteomic signatures: amino acid and oligopeptide compositions differentiate among phyla. *Proteins*, **54**, 20–40.

Petersen,T.N. *et al.* (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods*, **8**, 785–786.

Peterson,E.L. *et al.* (2009) Reduced amino acid alphabets exhibit an improved sensitivity and selectivity in fold assignment. *Bioinformatics*, **25**, 1356–1362.

Portugaly,E. *et al.* (2002) Selecting targets for structural determination by navigating in a graph of protein families. *Bioinformatics*, **18**, 899–907.

Prilusky,J. *et al.* (2005) FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics*, **21**, 3435–3438.

Radivojac,P. *et al.* (2013) A large-scale evaluation of computational protein function prediction. *Nat. Methods*, **10**, 221–227.

Rentzsch,R. and Orengo,C.A. (2009) Protein function prediction—the power of multiplicity. *Trends Biotechnol.*, **27**, 210–219.

Rost,B. *et al.* (2003) Automatic prediction of protein function. *Cell. Mol. Life Sci.*, **60**, 2637–2650.

Saeys,Y. *et al.* (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics*, **23**, 2507–2517.

Soding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.

Sonnhammer,E.L. *et al.* (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, **28**, 405–420.

Southey,B.R. *et al.* (2006) NeuroPred: a tool to predict cleavage sites in neuropeptide precursors and provide the masses of the resulting peptides. *Nucleic Acids Res.*, **34**, W267–W272.

Todd,A.E. *et al.* (2005) Progress of structural genomics initiatives: an analysis of solved target structures. *J. Mol. Biol.*, **348**, 1235–1260.

Vacic,V. *et al.* (2007) Composition Profiler: a tool for discovery and visualization of amino acid composition differences. *BMC Bioinformatics*, **8**, 211.

Valencia,A. (2005) Automatic annotation of protein function. *Curr. Opin. Struct. Biol.*, **15**, 267–274.

van den Berg,B.A. *et al.* (2014) SPiCE: a web-based tool for sequence-based protein classification and exploration. *BMC Bioinformatics*, **15**, 93.

Varshavsky,R. *et al.* (2007) When less is more: improving classification of protein families with a minimal set of global. In: Giancarlo,R. and Hannenhalli,S. (eds.) *Algorithms in Bioinformatics: 7th International Workshop, WABI*. Springer, Philadelphia, PA, pp. 12–24.

Veenstra,J.A. (2000) Mono- and dibasic proteolytic cleavage sites in insect neuroendocrine peptide precursors. *Arch. Insect Biochem. Physiol.*, **43**, 49–63.

Wang,L. *et al.* (2010) BindN + for accurate prediction of DNA and RNA-binding residues from protein sequence features. *BMC Syst. Biol.*, **4**(Suppl 1), S3.

Weathers,E.A. *et al.* (2004) Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein. *FEBS Lett.*, **576**, 348–352.

Wu,C.H. *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34**, D187–D191.

Yachdav,G. *et al.* (2014) PredictProtein—an open resource for online prediction of protein structural and functional features. *Nucleic Acids Res.*, **42**, W337–W343.

Zhang,G. and Fang,B. (2007) LogitBoost classifier for discriminating thermophilic and mesophilic proteins. *J. Biotechnol.*, **127**, 417–424.