Acta Biotheor (2018) 66:135–148
https://doi.org/10.1007/s10441-018-9327-x

REGULAR ARTICLE

# Mathematical Basis of Predicting Dominant Function in Protein Sequences by a Generic HMM–ANN Algorithm

**Siddhartha Kundu**[1,2]

Received: 18 April 2017 / Accepted: 16 April 2018 / Published online: 26 April 2018


**Abstract** The accurate annotation of an unknown protein sequence depends on extant data of template sequences. This could be empirical or sets of reference sequences, and provides an exhaustive pool of probable functions. Individual methods of predicting dominant function possess shortcomings such as varying degrees of inter-sequence redundancy, arbitrary domain inclusion thresholds, heterogeneous parameterization protocols, and ill-conditioned input channels. Here, I present a rigorous theoretical derivation of various steps of a generic algorithm that integrates and utilizes several statistical methods to predict the dominant function in unknown protein sequences. The accompanying mathematical proofs, interval definitions, analysis, and numerical computations presented are meant to offer insights not only into the specificity and accuracy of predictions, but also provide details of the operatic mechanisms involved in the integration and its ensuing rigor. The algorithm uses numerically modified raw hidden markov model scores of well defined sets of training sequences and clusters them on the basis of known function. The results are then fed into an artificial neural network, the predictions of which can be refined using the available data. This pipeline is trained recursively and can be used to discern the dominant principal function, and thereby, annotate an unknown protein sequence. Whilst, the approach is complex, the specificity of the final predictions can benefit laboratory workers design their experiments with greater confidence.

**Keywords** Algorithm · Artificial neural network · Dominant protein function · Hidden markov model · Subfamily

✉ Siddhartha Kundu
siddhartha_kundu@yahoo.co.in

[1] Department of Biochemistry, Dr. Baba Saheb Ambedkar Medical College and Hospital, Government of NCT of Delhi, Sector – 6, Rohini, Delhi 110085, India

[2] School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Mehrauli Road, New Delhi 110067, India

## 1 Background

The reliable annotation of genomic data is dependent on the assignment of function to protein sequences. Much of this information is gleaned from the clustering of these with existing functional groups. The presence of experimentally available data is invaluable to this effort, and in its absence the same has to be inferred from sequence data. This decomposition, into a superset of distinct functions of its constituent members (superfamily, family), is the most critical step of any clustering schema. A superfamily, by definition consists of sequences with poor, if any, sequence identity, with the simultaneous presence of one or more common fold(s). Consider the enzymes that belong to the iron $\left(Fe^{2+}\right)$ and 2-oxoglutarate (2OG) or $\alpha$-ketoglutarate (AKG) dependent dioxygenases (EC 1.14.11.x). The average inter-sequence identity of these enzymes ($< 25\%$), notwithstanding, the unifying features of these enzymes are the presence of a jelly-roll motif (Double strand $\beta$-helix; DSBH), and the substrate hydroxylating triad of residues $\left(HX[DE]X_nH\right)$ (Clifton et al. 2006; Hausinger 2004; Koehntop et al. 2005). However, the chemical nature of the cognate substrate(s) of these enzymes and/or the reactions differs substantially, and can form smaller clusters (Kundu 2012, 2015). Similarly, whilst the glycoside hydrolases (GHs 1-130; EC 3.2.1.x), comprise the larger set, plant GH9 endoglucanases can be further stratified into classes A, B, and C (Libertini et al. 2004; Lombard et al. 2014; Molhoj et al. 2002; Urbanowicz et al. 2007).

Whilst, the spatial arrangement of atoms of members of a superfamily dictates their biological role, differential function in a family of sequences can be attributed to the presence (native, acquired) or absence (native, excised) of specific sequence segments (classes A, B, and C of the plant GH9 endoglucanase family) and/or a limited number of amino acid residues (desaturases, demethylases, and chlorinating enzymes of the 2OG dependent dioxygenase superfamily). These regions are rarely silent, and can influence the behavior of the protein product(s) in vivo. Thus, while enzyme catalysis is dependent on conserved amino acids that form its active site geometry, generic proteins possess protein–protein, DNA/RNA–protein, transmembrane (TM), localization signals, and protein anchor-membrane domains that can influence its function. Despite the significant reduction in the dimensions of the superset to these smaller clusters, the unambiguous assertion of dominant function, remains challenging. For example, prevailing literature suggests that class B GH9 endoglucanases are the dominant forms of this family, far exceeding class C enzymes; a finding that is based on similarity to a few reference sequences (Buchanan et al. 2012; Montanier et al. 2010; Xie et al. 2013). Quantitative analyses of the differences between catalytically relevant segments of these enzymes, however, suggests that putative class C enzymes may approximate those of class B members (Kundu and Sharma 2016).

Sequence based classifiers of protein function can either be direct and deploy hidden markov models (HMMs), support vector machines (SVMs), and artificial neural networks (ANNs). Indirect indices of function range from domain comparison against existing databases such as the conserved domain database (CDD) of the national center for biotechnology information (NCBI), and the prediction of

secondary structural elements (Cao et al. 2016; Frishman and Argos 1995; Kabsch and Sander 1983; Marchler-Bauer et al. 2015; Martin et al. 2005). SVMs, although exhaustive, mandates the presence of training sets with highly similar sequences (Cao et al. 2016; Frishman and Argos 1995; Martin et al. 2005). Profile HMMs (pHMMs), are global representations of a multiple sequence alignment (MSA), and encompass modular information using a system of threshold values. A major finding in work done previously, however, highlighted the insensitivity of the inclusion thresholds, despite, log-orders of difference in the E-values used (Kundu and Sharma 2016). ANNs, are weighted approximations of multiple inputs to a function, and introduce bias in their computations as a means of achieving convergence. The reduction, to a single output channel, implies that this value is intrinsically ill-conditioned with the final prediction depending on the quality of the input. The arguments vide supra, justify the use of multiple statistical methods to assign dominant function to a protein of uncertain function. A specific instance (prediction of enzyme catalysis) of this pipeline has been tested on available sequence data in sequenced green plants (Kundu and Sharma 2016).

The work presented here is a detailed exposition of the mathematics that underlies the observed specificity and accuracy of a generic HMM–ANN algorithm in predicting dominant probable function in an unknown protein sequence. Detailed proofs for all the steps and the derivation of the unique intervals both, theoretical and observed that encompass the ANN predictions are presented and are meant to offer mechanistic insights into the process of integrating several statistical methods as well as the rigor that may ensue. In addition, the definition, analysis, and the numerical computation of bounds of the participating sets and intervals are discussed in context of selecting suitable datasets and dictating the architecture of the ANN deployed. Additionally, interesting mathematical results based on the Lebesgue outer measure are discussed along with its biological relevance.

## 2 Algorithm and Results

Step 0 Data collation, pre-processing, and computational tools. Protein sequences with detailed and specific biochemical data (kinetics, structure, mutagenesis) are preferred for training the HMMs and the ANN, while the test dataset can comprise sequences with expression data, unannotated coding segments of sequenced genomes (open reading frames, ORFs), or sequences with putative function. An alignment generating tool (Structural Alignment of Proteins, STRAP; Clustal suite), and HMMER (downloadable or server-based) may be used for model building, analysis, database construction, and similarity studies (Finn et al. 2015; Gille et al. 2014; Sievers and Higgins 2014). A scripting language (R, PERL, Python, AWK) may be utilized to analyze the data and perform miscellaneous tasks such as tabulation and formatting. The specialized R-packages needed to implement the unsupervised (clustering; *cluster*, *fpc*) and

supervised (ANN; *nnet*, *neuralnet*) machine learning tools utilized by this algorithm can be easily downloaded.

Step 1 Define and delineate the functions ($1 \leq \min(n) \leq n; n \in \mathbb{N}$) that an arbitrary protein sequence may be partitioned into. Utilize the clustering schema, i.e., primary ($\boldsymbol{A}$), secondary ($\boldsymbol{B}$), and tertiary ($\boldsymbol{D}$), to group the raw HMM-scores $\left(\alpha; \alpha \in \mathbb{R}_+, \mathcal{N}(0,1)\right)$. Whilst, the lower bounds of these ($\min(n) = min|\boldsymbol{A}| = min|\boldsymbol{B}| = min|\boldsymbol{D}| = 3$) are axiomatic (Defs. 1–3), the upper bounds may be inferred (Eqs. 1–4) (Table 1, Fig. 1b, c, d) (Kundu 2017). Briefly,

$$\boldsymbol{A} = \left\{\alpha_i | \alpha \in \mathbb{R}_+, \mathcal{N}(0,1); 1 < i \leq \min(n); \quad i, n \in \mathbb{N}\right\} \quad \text{(Def. 1)}$$

$$\boldsymbol{B} = \left\{\left(\alpha_i, \alpha_j\right) | \alpha \in \mathbb{R}_+, \mathcal{N}(0,1); 1 < i, j \leq \min(n); \quad i \neq j, j, n \in \mathbb{N}\right\} \quad \text{(Def. 2)}$$

$$\boldsymbol{D} = \left\{\left(\left(\alpha_i, \alpha_j\right), \left(\alpha_j, \alpha_k\right)\right) | \alpha \in \mathbb{R}_+, \mathcal{N}(0,1); 1 < i, j, k \leq \min(n); \quad i \neq j \neq k; \quad i, j, k, n \in \mathbb{N}\right\} \quad \text{(Def. 3)}$$

**Table 1** Role of inputs in defining ANN architecture

| \|A\| | \|B\| | \|D\| | H1 | H2 | H3 |
|---|---|---|---|---|---|
| 3 | 3 | 3 | 4 | 4 | 3 |
| 4 | 6 | 15 | 12 | 8 | 11 |
| 5 | 10 | 45 | 30 | 14 | 31 |
| 6 | 15 | 105 | 63 | 21 | 71 |
| 7 | 21 | 210 | 120 | 29 | 141 |
| 8 | 28 | 378 | 209 | 39 | 253 |
| 9 | 36 | 630 | 341 | 50 | 421 |
| 10 | 45 | 990 | 527 | 63 | 661 |
| 11 | 55 | 1485 | 782 | 77 | 991 |
| 12 | 66 | 2145 | 1119 | 93 | 1431 |
| 13 | 78 | 3003 | 1557 | 110 | 2003 |
| 14 | 91 | 4095 | 2112 | 128 | 2731 |
| 15 | 105 | 5460 | 2804 | 148 | 3641 |
| 16 | 120 | 7140 | 3655 | 169 | 4761 |
| 17 | 136 | 9180 | 4686 | 192 | 6121 |
| 18 | 153 | 11,628 | 5922 | 216 | 7753 |

**A**: Set of raw HMM scores of a protein sequence

**B**: Set of pairs of raw HMM scores of a protein sequence

**D**: Set of pairs-of-pairs of raw HMM scores of a protein sequence

**H1**: $0.5 * (|\boldsymbol{D}| + 1) + \sqrt{|\boldsymbol{D}|}$

**H2**: $2 * \sqrt{|\boldsymbol{D}| + 1}$

**H3**: $2 * (|\boldsymbol{D}| + 1)/3$

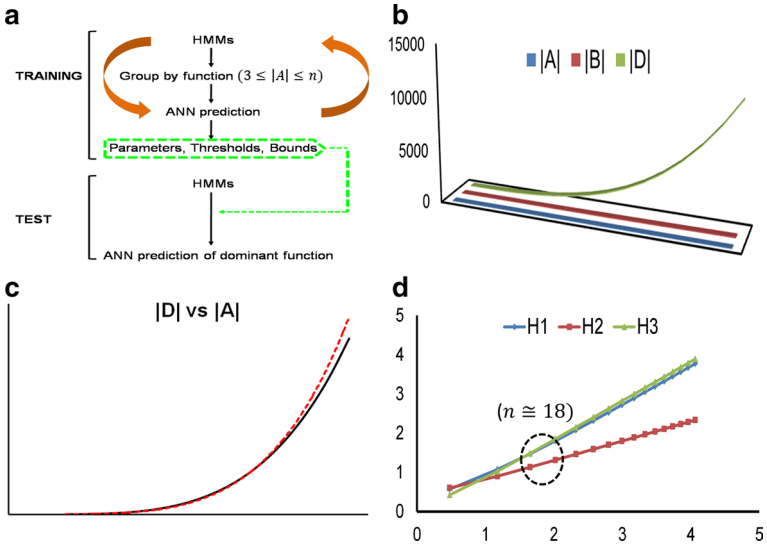


**Fig. 1** Generic algorithm for predicting dominant function in a protein sequence. **a** Steps needed to construct and validate the HMM–ANN algorithm on a well characterized training set. The datasets may be repeatedly sampled for parameter definition and model refinement. The final output is a set of high confidence bounds that is mapped and specific for each predicted function, **b** analysis of cardinality of various sets used in parameterization, **c** scatter plot between the number of predicted function and the pairs-of-pairs of modified HMM scores, and **d** relevance of cardinality of the superset of probable functions to the architecture of the ANN. Abbreviations: HMM, hidden markov model; ANN, artificial neural network; **A**, **B**, **D**, Sets of raw HMM scores; H1, H2, H3, Methods to compute number of nodes in the hidden layer of a 1:1:1 ANN

$$y_B = (0.5)\left(x_A^2\right) - (0.25)\left(x_A\right) + 2E - 13; \quad R^2 = 1.00 \tag{1}$$

$$y_B = (0.298)\left(x_A^{2.1704}\right); \quad R^2 = 0.9994 \tag{2}$$

$$y_D = (0.125)\left(x_A^4\right) - (0.25)\left(x_A^3\right) - (0.125)\left(x_A^2\right) + (0.25)\left(x_A\right) - 2E - 08; \quad R^2 = 1.00 \tag{3}$$

$$y_D = (0.0293)\left(x_A^{4.4996}\right); \quad R^2 = 0.9978 \tag{4}$$

Step 2 Define and enumerate the list of full length sequences that best represents each of these functions. These sequences $\{m \in \boldsymbol{G} | m \in \mathbb{N}\}$, then constitute the training dataset for each predicted function ($1 \leq \min(n) \leq n$) and must necessarily possess the recommended sequence suitability index ($SSI > 1.00$) (Kundu 2017). These could also be complemented with extant empirical data.

Step 3 Estimate the $\beta$-value $\left(\sum_{l=1}^{l=|\boldsymbol{D}|}\zeta_{nml}\right)$ for each sequence $\left(\beta_{nm};1\le \min(n)\le n, 1\le m\le |\boldsymbol{G}|\right)$ (Fig. 1a) (Kundu 2017; Kundu and Sharma 2016).

$\zeta$, Computed value of pairs-of-pairs of raw HMM scores of the $m$th sequence of the $n$th function; $\boldsymbol{D}$, Composite set of pairs-of-pairs of raw HMM scores of the $m$th sequence of the $n$th function; $n$, $m$th member of $n$th function; $m$, $m$th member of $n$th function; $l$, $l$th member of $\boldsymbol{D}$.

**Lemma** *The computed value* $\left(\zeta_{nml}\right)$ *of a pair-of-pairs* (*POP*) *of raw HMM scores is numerically equivalent to its z-score, i.e.,* $\zeta_{nml}\simeq z_{nml}$

*Proof* Define $\zeta_{nml}\left(\mu_\alpha,\sigma_\alpha\right)$ such that $\alpha\in\mathbb{R}_+,\mathcal{N}(0,1)$

In the absence of an explicit assumption of a normal population, the mean ($\mu$) and standard deviation ($\sigma$) are not independent, i.e., $\mu_\alpha\propto\sigma_\alpha$,

It follows that $\exists\left\{\zeta_{nml}\right\}_{l=1}^{l=|\boldsymbol{D}|};\zeta_{nml}\in\mathbb{R}_+,\mathcal{N}(0,1)$

$$\text{Then, } z_{nml}=\left(\zeta_{nml}-\bar{\zeta}_{nml}\right)/\sigma_{\zeta_{nml}}=\zeta_{nml}/\sigma_{\zeta_{nml}}-\bar{\zeta}_{nml}/\sigma_{\zeta_{nml}}=\zeta_{nml} \tag{5}$$

Step 4 The $\beta_{nm}$-values computed in Step 3 are then clustered, such that every cluster mean represents the centroid of a specific function $\left(\left\{\beta_n^{'}\right\};1\le\min(n)\le n;n\in\mathbb{N}\right)$. These are then compared $\left(\chi^2(n)=\sum_{m=1}^{\mathrm{m}=|\mathbf{G}|}\left(\beta_{nm}-\beta_n^{'}\right)^2/\beta_n^{'};\quad 1\le\min(n)\le n;\quad 1\le m\le|\boldsymbol{G}|;n,m\in\mathbb{N}\right)$

Since, the cluster means are derived from the sequence data their difference is expected to be trivial $\left(\min\left(\sum_{m=1}^{m=|\boldsymbol{G}|}\chi^2(n)\right)\right)$

subsume $\left(\chi^2(n)\to 0\right)$

$$\sum_{m=1}^{m=|G|}\left((\beta_{nm}-\beta_n^{'})^2/\beta_n^{'}\right)\to 0\,(1<\min(n)\le n;\quad 1\le m\le|\boldsymbol{G}|;\quad n,m\in\mathbb{N})$$

$$\left(\sum_{m=1}^{m=|\boldsymbol{G}|}(\beta_{nm}-\beta_n^{'})^2\right)/\beta_n^{'}\to 0$$

$$\left(\beta_{n1}^2+\left(\beta_n^{'}\right)^2-(2)\left(\beta_{n1}\right)\left(\beta_n^{'}\right)\right)+\left(\beta_{n2}^2+\left(\beta_n^{'}\right)^2-(2)\left(\beta_{n2}\right)\left(\beta_n^{'}\right)\right)/\beta_n^{'}\ldots\to 0$$

Rearranging the terms and differentiating w.r.t $\beta_n^{'}$

$$\left(\beta_{n1}^2+\beta_{n2}^2\cdots(|\boldsymbol{G}|)\left(\beta_n^{'}\right)^2-(2)\left(\beta_n^{'}\right)\left(\beta_{n1}+\beta_{n2}\cdots\right)\right)/\beta_n^{'}\to 0$$

$$\frac{d(\beta_{n1}^2)}{d\beta_n'} + \frac{d(\beta_{n2}^2)}{d\beta_n'} + \frac{(|\boldsymbol{G}|)d\left(\beta_n'\right)^2}{d\beta_n'} - \frac{(2)d\left(\beta_n'\beta_{n1} + \beta_n'\beta_{n2}\cdots\right)}{d\beta_n'} = 0 \tag{6}$$

$$(2)(|\boldsymbol{G}|)\left(\beta_n'\right) - 2\left(\beta_{n1} + \beta_{n2} + \cdots \beta_{nm}\right) = 0 \tag{7}$$

$$(2)(|\boldsymbol{G}|)\left(\beta_n'\right) - 2\sum_{m=1}^{m=|\boldsymbol{G}|}\beta_{nm} = 0$$

$$(|\boldsymbol{G}|)\left(\beta_n'\right) = \sum_{m=1}^{m=|\boldsymbol{G}|}\beta_{nm} = \left(\beta_{n1} + \beta_{n2} + \cdots \beta_{nm}\right) \tag{8}$$

Consider the arbitrary terms $\beta_{nm}, \beta_{n(m+i)} \forall i \neq m$
If $\beta_{n(m+i)} + \varepsilon > \beta_{nm}, \varepsilon \in \mathbb{R}_+$

$$\text{Then} \sum_{m=1}^{m=|\boldsymbol{G}|}\beta_{nm} + \varepsilon > (|\boldsymbol{G}|)\left(\beta_n'\right) \tag{9}$$

Similarly, if $\beta_{n(m+i)} - \varepsilon > \beta_{nm}, \varepsilon \in \mathbb{R}_+$

$$\text{Then} \sum_{m=1}^{m=|\boldsymbol{G}|}\beta_{nm} + \varepsilon < (|\boldsymbol{G}|)\left(\beta_n'\right) \tag{10}$$

From Eqs. (9) and (10),

$$\beta_{n(m+i)} = \beta_{nm} (\forall i \neq m)$$

$$\sum_{m=1}^{m=|\boldsymbol{G}|}\beta_{nm} = (|\boldsymbol{G}|)\left(\beta_{nm}\right) = (|\boldsymbol{G}|)\left(\beta_n'\right) \tag{11}$$

Substituting this value in (Eq. 7)

$$\left(\sum_{m=1}^{m=|\boldsymbol{G}|}(\beta_{nm} - \beta_{nm})/|\boldsymbol{G}|)^2\right)/\beta_n' = 0 \tag{12}$$

$$\beta_n' \simeq \beta_{nm}\left(\forall \beta_{nm}\right) \tag{13}$$

Step 5 Utilize the results in Step 4 in association with pre-computed values of the set of pairs-of-pairs for each sequence of each probable function ($\zeta_{nml}$; $1 < \min(n) \leq n$; $1 \leq m \leq |\boldsymbol{G}|$; $1 \leq l \leq |\boldsymbol{D}|$; $n, m, l \in \mathbb{N}$) and define the input $\left(\beta'\right)$ and output $\left(\beta''\right)$ channels to the artificial neural network (ANN) (Kundu and Sharma 2016):

$$\beta_n^{'} \simeq \beta_{nm} \simeq \beta_{nm}^{'} \simeq \sum_{l=1}^{l=|\boldsymbol{D}|} \zeta_{nml}$$

$$\begin{aligned} \beta_{nm}^{'} &\simeq \sum_{l=1}^{l=|\boldsymbol{D}|} \zeta_{nml} \\ &= \sum_{l=1}^{l=|\boldsymbol{D}|} \left(\lambda_{nml}\right)\left(\zeta_{nml}\right) \end{aligned} \tag{14}$$

$$= \beta_{nm}^{''} \tag{15}$$

$\zeta$, Computed value of pairs-of-pairs of raw HMM scores of the $m$th sequence of the $n$th function; $\lambda$, Weighted $\zeta$-score computed by the ANN; $\boldsymbol{D}$, Composite set of pairs-of-pairs of raw HMM scores of the $m$th sequence of the $n$th function.

Step 6 Define the intervals $\left(\mathcal{I}_n\right)$ unique to each probable set of functions that an unknown protein sequence may be assigned to. These could be estimated directly or determined empirically $\left(prediction \rightarrow (\beta_{nm}^{''} \pm \varepsilon) \wedge \zeta_{nml}\right.$; $\zeta, \varepsilon \in \mathbb{R}_+$; $1 < \min(n) \leq n$; $1 \leq m \leq |\boldsymbol{G}|$;$1 \leq l \leq |\boldsymbol{D}|$;$n, m, l \in \mathbb{N}$) (Def. 3) (Kundu and Sharma 2016).

$$\mathcal{I}_n = \begin{cases} \beta_n^{'} \pm \left| \left(t_{\alpha/2}\right)\left(\sigma/\sqrt{|\boldsymbol{G}|}\right) \right|, & |\boldsymbol{G}| < 30 \\ \beta_n^{'} \pm \left| (z)\left(\sigma/\sqrt{|\boldsymbol{G}|}\right) \right|, & |\boldsymbol{G}| \geq 30 \end{cases} \tag{16}$$

$\beta_n^{'}$, Centroid of $n$th cluster; $t_{\alpha/2}$, Interval coefficient of upper tail of $t$-distribution; $z$, Interval coefficient of normal distribution; $\sigma$, Standard deviation of sample; $|\boldsymbol{G}|$, Size of $n$th cluster; $m$, $m$th *member of* $n$th *cluster.*

Step 7 Define the bounds $(a, b)$ of the search space by considering the countable union of the sequence of open and pairwise disjoint intervals (observed, expected) contained within the encompassing major interval $\left(\boldsymbol{\mathcal{J}}_{[a,b]} = \bigcup_{n=1}^{n \geq \min(n)} \beta_n^{'}; \beta_n^{'} \in (\beta_n^{'} - \sigma_n, \beta_n^{'} + \sigma_n);\quad 1 < \min(n) \leq n\right)$. The

size $\left(l\left(\boldsymbol{\mathcal{J}}_{[a,b]}\right)\right)$ is then the outer Lebesgue measure $m^*\left(\boldsymbol{\mathcal{J}}_{[a,b]}\right)$ of the encompassing interval.

$$l\left(\boldsymbol{\mathcal{J}}_{[a,b]}\right) = m^*\left(\boldsymbol{\mathcal{J}}_{[a,b]}\right) = |b-a| \tag{17}$$

$b$, max $\left(\beta_n^{'}\right) + \sigma_n$; $a$, min $\left(\beta_n^{'}\right) - \sigma_n$; $\sigma_n$, Standard deviation of $n$th cluster; $\beta_n^{'}$, Centroid of $n$th cluster.

Step 8 Validate ANN-predictions $\left(\beta''\right)$ of dominant function for the training sequences. This could be: (a) an exhaustive cross validation of each sequence of each probable function ($|\boldsymbol{G}| < 30$), (b) performed on a distinct validation subset ($\approx 25-30\%$) of the training sequences if the sample sizes are adequate ($|\boldsymbol{G}| \geq 30$), or (c) empirical using pre-defined criteria appropriate to the dataset examined such as ($\beta''_{nm} \cong \beta'_{nm} := max\ (HMM); 1 < \min\ (n) \leq n, 1 \leq m \leq |\boldsymbol{G}|$) (Kundu and Sharma 2016).

## 3 Discussion

### 3.1 Contribution of the Probability of Mapping the ANN-Prediction to a Distinct Partition

The effective prediction by the ANN of dominant function for an unknown protein sequence $\left(\beta''_{seq} \in \mathbb{R}_+\right)$ is dependent on it being unambiguously mapped to a single numerical interval whose centroids approximate the cluster means for that particular function. Consider the closed and bounded interval of length $\left(l\left(\boldsymbol{\mathcal{J}}_{[a,b]}\right) = |b-a|\right)$ (Step 7; Eq. 17) and the following sequences of open and pairwise disjoint subintervals (Step 6):

Consider the sequence ($\boldsymbol{\mathcal{H}} \subseteq \boldsymbol{\mathcal{J}}$) of uniquely observed open and pairwise disjoint subintervals:

$$\boldsymbol{\mathcal{H}} = \bigcup_{n=1}^{\mathrm{n}\geq\min(n)} \left(\beta_n^{'} - \sigma_n, \beta_n^{'} + \sigma_n\right)$$

$$m^*(\boldsymbol{\mathcal{H}}) = m^*\left(\bigcup_{n=1}^{\mathrm{n}\geq\min(n)} \left(\beta_n^{'} - \sigma_n, \beta_n^{'} + \sigma_n\right)\right)$$
$$= \sum_{n=1}^{\mathrm{n}\geq\min(n)} l(\beta_n^{'} - \sigma_n, \beta_n^{'} + \sigma_n) = \sum_{n=1}^{\mathrm{n}\geq\min(n)} l(\phi) = \{0\}_{n=1}^{\mathrm{n}\geq\min(n)}$$

Consider the covering of sequences of arbitrary open and pairwise disjoint intervals ($\boldsymbol{\mathcal{L}} \subseteq \boldsymbol{\mathcal{J}}$)

$$\boldsymbol{\mathcal{L}} = \bigcup_{p=1}^{p=P} \left(a_p, b_p\right); a, b \in \{\mathbb{Z}_+, \mathbb{R}_+\}$$

$$\exists q_p \in \left(a_p, b_p\right); q_p \in \mathbb{Q}; \because \mathbb{Q} \textit{ is dense in } \mathbb{R}$$

$$a_p < q_p < b_p$$

$$a_p < q_p \Rightarrow a_p + \varepsilon = q_p \textit{ or } a_p = q_p - \varepsilon; \quad \varepsilon \in \mathbb{R}_+ \tag{18}$$

$$b_p > q_p \Rightarrow b_p - \varepsilon = q_p \textit{ or } b_p = q_p + \varepsilon; \quad \varepsilon \in \mathbb{R}_+ \tag{19}$$

$$q_p - \varepsilon < q_p < q_p + \varepsilon; \quad \varepsilon \in \mathbb{R}_+ (\text{Eqs. } 18, 19)$$

$$\Rightarrow q_p \in (q_p - \varepsilon, q_p + \varepsilon)$$

$$\text{Similarly, } \beta_n^{'} \in (\beta_n^{'} - \sigma, \beta_n^{'} + \sigma)$$

$$q_p = \beta_n^{'}; \textit{ iff } \varepsilon = \sigma; \varepsilon, \sigma \in \mathbb{R}_+ \tag{20}$$

Rewriting,

$$\boldsymbol{\mathcal{L}} = \bigcup_{p=1}^{p=P} \left(q_p - \varepsilon, q_p + \varepsilon\right)$$

$$m^*(\boldsymbol{\mathcal{L}}) = m^*\left(\bigcup_{p=1}^{\mathrm{P}} \left(q_p - \varepsilon, q_p + \varepsilon\right)\right) = \sum_{p=1}^{\mathrm{P}} l\left(q_p - \varepsilon, q_p + \varepsilon\right) = \sum_{p=1}^{\mathrm{P}} l(\phi) = \{0\}_{p=1}^{\mathrm{P}}$$

$$m^*(\boldsymbol{\mathcal{L}}) = m^*(\boldsymbol{\mathcal{H}}) = 0 \tag{21}$$

Despite the result in Eq. 21, $|\boldsymbol{H}| \leq |\boldsymbol{\mathcal{L}}|$ and as $P \to \infty$, $|\boldsymbol{H}| \lll |\boldsymbol{\mathcal{L}}|$.

The probability of mapping each ANN-output $\left(\beta''\right)$ to a distinct sub-partition ($\tau = 1/|\boldsymbol{\mathcal{H}}| * |\boldsymbol{\mathcal{L}}| \simeq 1/|\boldsymbol{\mathcal{H}}|$) (Eq. 22).

**Theorem** *The number of probable functions* $|\mathbf{A}|$ *for any defined interval is countably infinite.*

*Proof* Consider the aforementioned sets $(\boldsymbol{\mathcal{H}}, \boldsymbol{\mathcal{L}})$. Since, every probable function is modeled as an open and bounded interval with a centroid, and $\mathbb{Q}$ is dense in $\mathbb{R}$, we can always find an infinite number of rational numbers between any two real numbers, *i.e.*,

$$a_p - q_p < 1/x \Longrightarrow a_p < 1/x + q_p$$

$$1/y < b_p - q_p \Longrightarrow 1/y + q_p < b_p$$

Rewriting these inequalities and continuing,

$$a_p < q_p < 1/x + q_p + \cdots \leq \cdots 1/y + q_p < b_p \tag{22}$$

$$a_p, b_p \in \mathbb{R}_+ (\textit{Set of positive real numbers})$$
$$q_p, 1/x, 1/y \in \mathbb{Q}\,(\textit{Set of rational numbers})$$
$$x, y \in \mathbb{Z}_+ (\textit{Set of positive integers})$$

### 3.2 Relevance of Functional Constraints to Unambiguous Assignment of Dominant Function

The outlined protocol is expected to improve upon previous stratification attempts, both, in terms of biological relevance, as well as in the accuracy of predictions. The latter has been assessed in earlier work using the indices of precision (specificity) and recall (sensitivity) (Kundu and Sharma 2016). Whilst, the utility of collating biochemical data relevant to sequence clustering is unequivocal; the multitude of methods utilized imposes rigor in the schema. In particular, the use of the SSI (Step 2) in tandem with empirical data can refine the selection of training sequences such that $\beta$-value for each relevant sequence $\left(\beta_{nm}\right)$ is within one standard deviation of the centroid for a particular cluster $(|\beta_{nm} - \beta_n^{'}| < \sigma_n)$ and may even converge $(|\beta_{nm} - \beta_n^{'}| \to 0)$ (Steps 3 and 4) (Kundu 2017). The Chi squared data (Step 4), too, can be utilized to modify this selection such that an outlier sequence can be edited at this stage as well. The ratio of the input and output channels is critical to accomplishing convergence in an ANN (Step 5) with multiple outputs, as is its determination of the number of hidden layers. In contrast, despite a single output's risk at being ill-conditioned, the unbiased assignment of dominant function mandates its persistent use. Clearly, well partitioned (open, bounded, pairwise disjoint) intervals that encompass the inputs $\left(\zeta_{nml}\right)$ to the ANN are then a pre-requisite for efficacious prediction (Steps 6 and 7). The number of theoretical partitions $\left(\zeta_{nml} \in \boldsymbol{\mathcal{L}}; \zeta_{nml} \to \infty\right)$ (Steps 6 and 7), notwithstanding, the analysis suggests that the cardinality of the superset $(|\boldsymbol{A}|)$ of probable functions that an unknown sequence may be partitioned into is important and must be considered (Step 1).

Consider the following functions $\left(f : \boldsymbol{A} \to \{|\boldsymbol{A}|\}_{k=3}^{K}; g : \boldsymbol{D} \to \{|\boldsymbol{D}|\}_{k=3}^{K}\right)$ (Step 1) (Table 1, Fig. 1c)

Clearly, $f(\boldsymbol{A}) \sim g(\boldsymbol{D})(K \to \infty)$ and $1/f(\boldsymbol{A})g(\boldsymbol{D}) \simeq \frac{1}{f(\boldsymbol{A})}$

$$^{1}/_{f(\boldsymbol{A})} = \tau \tag{23}$$

$\boldsymbol{A}$, Raw HMM scores of all probable functions for a sequence; $\boldsymbol{D}$, Composite set of pairs-of-pairs of raw HMM scores of the $m$th sequence of the $n$th function; $\tau$, Probability of assigning a unique dominant function to a protein sequence.

Prediction of dominant function by this integrated algorithm is also likely to be constrained at the ANN stage, wherein, a larger number of hidden neurons may not result in any additional information. Extant literature from clinical medicine, agriculture, and academia, that have utilized ANN-based predictors suggests that the upper limit for neurons/nodes in a back-propagation (BP) ANN with 1 : 1 : 1 architecture is $n \cong 18$ (Akbari Hasanjani and Sohrabi 2017; Kundu and Sharma 2016; Hawari and Alnahhal 2016; Teshnizi and Ayatollahi 2015; Shi et al. 2013; Yamamura et al. 2008; Zhou and Li 2007). This, in turn would imply a limit on the cardinality of the superset of all probable functions that a protein sequence might be expected to possess, i.e., $3 \leq |\boldsymbol{A}| \leq 6$; $0.166 \leq \tau \leq 0.33$ (Table 1, Fig. 1d).

## 4 Concluding Remarks

The HMM–ANN based algorithm accurately predicts dominant biological function of an unknown protein sequence. The detailed mathematical treatment of the various steps of this algorithm not only offers insights into the origins of this specificity, but also highlights the mechanism of integrating multiple methods into a generic functional algorithm. Additionally, it may assist investigators in preparing a computationally feasible superset, of putative function for their sequence(s) of interest. The algorithm itself, can be adapted with little effort, and uses publically available software and tools. The coding, when needed is trivial and can be accomplished with ease. The computations are self explanatory, lucid, and can be readily comprehended by biologists. The existence of upper and lower bounds may impose constraints on the selection of features/probable functions that could characterize a protein sequence. However, careful curation, inclusion of empirical data, and strict thresholds could go a long way in broadening the utility of this generic HMM–ANN algorithm.

**Author's Contribution** SK formulated, designed, and wrote the algorithm, developed and tested the filters, carried out the mathematical and computational analysis, wrote all necessary code, and the manuscript.

**Compliance with Ethical Standards**

**Conflict of interest** The authors declare that they have no competing interests.

## References

Akbari Hasanjani HR, Sohrabi MR (2017) Artificial neural networks (ANN) for the simultaneous spectrophotometric determination of fluoxetine and sertraline in pharmaceutical formulations and biological fluid. Iran J Pharm Res 16:478–489

Buchanan M, Burton RA, Dhugga KS, Rafalski AJ, Tingey SV, Shirley NJ, Fincher GB (2012) Endo-(1,4)-beta-glucanase gene families in the grasses: temporal and spatial co-transcription of orthologous genes. BMC Plant Biol 12:235

Cao C, Wang G, Liu A, Xu S, Wang L, Zou S (2016) A new secondary structure assignment algorithm using calpha backbone fragments. Int J Mol Sci 17(3):333

Clifton IJ, McDonough MA, Ehrismann D, Kershaw NJ, Granatino N, Schofield CJ (2006) Structural studies on 2-oxoglutarate oxygenases and related double-stranded beta-helix fold proteins. J Inorg Biochem 100(4):644–669

Finn RD, Clements J, Arndt W, Miller BL, Wheeler TJ, Schreiber F, Bateman A, Eddy SR (2015) HMMER web server: 2015 update. Nucleic Acids Res 43(W1):W30–W38

Frishman D, Argos P (1995) Knowledge-based protein secondary structure assignment. Proteins 23(4):566–579

Gille C, Fahling M, Weyand B, Wieland T, Gille A (2014) Alignment-Annotator web server: rendering and annotating sequence alignments. Nucleic Acids Res 42(Web Server issue):W3–W6

Hausinger RP (2004) FeII/alpha-ketoglutarate-dependent hydroxylases and related enzymes. Crit Rev Biochem Mol Biol 39(1):21–68

Hawari AH, Alnahhal W (2016) Predicting the performance of multi-media filters using artificial neural networks. Water Sci Technol 74:2225–2233

Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22(12):2577–2637

Koehntop KD, Emerson JP, Que L Jr (2005) The 2-His-1-carboxylate facial triad: a versatile platform for dioxygen activation by mononuclear non-heme iron(II) enzymes. J Biol Inorg Chem 10(2):87–93

Kundu S (2012) Distribution and prediction of catalytic domains in 2-oxoglutarate dependent dioxygenases. BMC Res Notes 5:410

Kundu S (2015) Unity in diversity, a systems approach to regulating plant cell physiology by 2-oxoglutarate-dependent dioxygenases. Front Plant Sci 6:98

Kundu S (2017) Mathematical basis of improved protein subfamily classification by a HMM-based sequence filter. Math Biosci 293:75–80

Kundu S, Sharma R (2016) In silico identification and taxonomic distribution of plant class C GH9 endoglucanases. Front Plant Sci 7(1185):1–21

Libertini E, Li Y, McQueen-Mason SJ (2004) Phylogenetic analysis of the plant endo-beta-1,4-glucanase gene family. J Mol Evol 58(5):506–515

Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B (2014) The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res 42(Database issue):D490–D495

Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, Lanczycki CJ, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D, Zheng C, Bryant SH (2015) CDD: NCBI's conserved domain database. Nucleic Acids Res 43(Database issue):D222–D226

Martin J, Letellier G, Marin A, Taly JF, de Brevern AG, Gibrat JF (2005) Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. BMC Struct Biol 5:17

Molhoj M, Pagant S, Hofte H (2002) Towards understanding the role of membrane-bound endo-beta-1,4-glucanases in cellulose biosynthesis. Plant Cell Physiol 43(12):1399–1406

Montanier C, Flint JE, Bolam DN, Xie H, Liu Z, Rogowski A, Weiner DP, Ratnaparkhe S, Nurizzo D, Roberts SM, Turkenburg JP, Davies GJ, Gilbert HJ (2010) Circular permutation provides an evolutionary link between two families of calcium-dependent carbohydrate binding modules. J Biol Chem 285(41):31742–31754

Shi L, Wang XC, Wang YS (2013) Artificial neural network models for predicting 1-year mortality in elderly patients with intertrochanteric fractures in China. Braz J Med Biol Res 46:993–999

Sievers F, Higgins DG (2014) Clustal omega, accurate alignment of very large numbers of sequences. Methods Mol Biol 1079:105–116

Teshnizi SH, Ayatollahi SM (2015) A comparison of logistic regression model and artificial neural networks in predicting of Student's Academic failure. Acta Inform Med 23:296–300

Urbanowicz BR, Bennett AB, Del Campillo E, Catala C, Hayashi T, Henrissat B, Hofte H, McQueen-Mason SJ, Patterson SE, Shoseyov O, Teeri TT, Rose JK (2007) Structural organization and a standardized nomenclature for plant endo-1,4-beta-glucanases (cellulases) of glycosyl hydrolase family 9. Plant Physiol 144(4):1693–1696

Xie G, Yang B, Xu Z, Li F, Guo K, Zhang M, Wang L, Zou W, Wang Y, Peng L (2013) Global identification of multiple OsGH9 family members and their involvement in cellulose crystallinity modification in rice. PLoS ONE 8(1):e50171

Yamamura S, Kawada K, Takehira R, Nishizawa K, Katayama S, Hirano M, Momose Y (2008) Prediction of aminoglycoside response against methicillin-resistant *Staphylococcus aureus* infection in burn patients by artificial neural network modeling. Biomed Pharmacother 62:53–58

Zhou R, Li Y (2007) Texture analysis of MR image for predicting the firmness of Huanghua pears (Pyrus pyrifolia Nakai, cv. Huanghua) during storage using an artificial neural network. Magn Reson Imaging 25:727–732