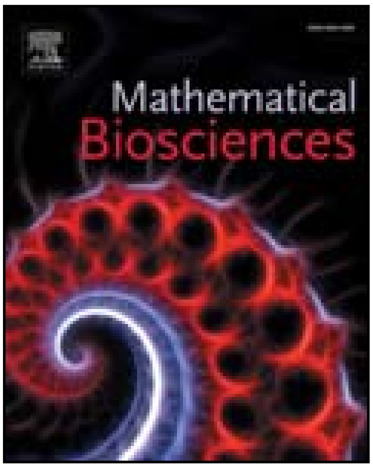
Accepted Manuscript



Mathematical basis of improved protein subfamily classification by a HMM-based sequence filter

Siddhartha Kundu

PII: S0025-5564(16)30301-7
DOI: 10.1016/j.mbs.2017.09.001
Reference: MBS 7973

To appear in: *Mathematical Biosciences*

Received date: 9 November 2016
Revised date: 14 June 2017
Accepted date: 11 September 2017

Please cite this article as: Siddhartha Kundu , Mathematical basis of improved protein subfamily classification by a HMM-based sequence filter, *Mathematical Biosciences* (2017), doi: 10.1016/j.mbs.2017.09.001

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

INTRODUCTION

Any meaningful classification of protein sequences needs to be based on analogy and represents, the most critical step in any clustering schema. In its absence, protein sequence classification methods are dependent on the accuracy and criteria used in the collation process. Phylogenetic analysis utilizes information gleaned from structural or sequence based alignments to infer molecular ancestry, divergence, and subfamily membership. Whilst, sequence identity and/ similarity remain the major tenet of most taxonomic grouping schemas, combination regimens attempt to incorporate elements of higher order protein structure. These include secondary structural elements, domains, folds (protein specific), as well as functional data (substrate interacting active site residues) (Jones 1999a; Jones 1999b; Kundu 2012). The improvements in algorithms and related methods have resulted in a continuum of progressively more accurate and relevant classification protocols. Despite these developments, high throughput sequencing with its voluminous output of data results in the emergence of novel sequences with varying degrees of inter-molecular similarity. An interesting subset are those which possess above average sequence similarity, which in tandem with a purported spectra of activity can influence the robustness, and thereby, the confidence of any clustering protocol. Despite the paucity of corroborating laboratory data, there is a need to define high quality clusters of sequences which can be utilized as molecular templates.

Differential activity in association with protein sequence conservation is not unknown in paralogs. These sequences, despite, their common molecular evolutionary ancestry have diverged, with the incorporation and/or loss of amino acids or segments of sequences critical to function. A suitable definition for these sequences whence subjected to the above constraints, is that of a confounding sequence. Formally, a sequence with moderate/ high amino acid identity with representative members of several groups, and whose multiplicity of probable functions impedes its unique clustering, can be considered as such. Thus, plant GH9 endoglucanases with class C activity possess an additional and unique substrate modifying domain. This carbohydrate binding module (CBM49) enables the digestion, by plant enzymes of microcrystalline regions of cellulose, a property shared with some bacteria (Chung et al. 2015; Urbanowicz et al. 2007). Nevertheless, these enzymes share demonstrable sequence identity with class B enzymes

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

$(30 - 50\%)$, a feature that has given rise to considerable debate on classifying these enzymes (Kundu and Sharma 2016). Similarly, YLL0577c, a putative Fe (II) and alpha-ketoglutarate dependent dioxygenase from *S. cerevisiae* shares $(\approx 30 - 35\%)$ similarity, which includes the conservation of a modified active site geometry $((HX[DE]X_nH) \rightarrow (HX_{20-30}HX[DE]X_nH))$, with the bonafide sulfur metabolizing taurine dioxygenase and alkylsulfatase enzymes (Hogan et al. 1999). Additionally, however, there is a $(\approx 25 - 30\%)$ similarity with the 2,4-dichlorophenoxyacetic acid (2, 4-D) degrading enzyme (tfdA) (Fukumori and Hausinger 1993; Hogan et al. 1999). Kinetic and mutagenesis studies clearly suggest a preference for sulfonates such as taurocholate and isethionate (Hogan et al. 1999). These examples highlight not only the importance of supporting biochemical data in predicting function, but, also the ambiguity that could result in assigning membership due to its lack, thereof.

Methods of evaluating confounding sequences range from single to integrated numerical protocols. If the inter-sequence identity/ similarity is high $(51 - 75\%)$, as in members of a protein subfamily, percent accepted mutations (PAM) in association with support vector machines (SVMs) can be utilized to map the progressive divergence of function (drift), to the sequence(s) of interest (Khater and Mohanty 2015; Mount 2008). On the other hand, for sequences with moderate identity/ similarity $(25 - 50\%)$, the predicted loss- or gain-of function (shift) can be mapped to the presence/ absence of specific sequence segments. This mandates the usage of matrices such as the blocks amino acid substitution matrices (BLOSUM), which, along with comprehensive stochastic methods such as Hidden Markov Models (HMMs) of the resulting alignment (Gene3D, SMART, Pfam, H2OGpred, DB2OG) can offer insights into putative function (Kundu 2012; Kundu 2015; Lees et al. 2010; Letunic et al. 2002; Mount 2008; Sonnhammer et al. 1998). The generic properties of confounding sequences and their influence on the accuracy of predictions has been discussed previously (Kundu and Sharma 2016). A notable finding, of this analysis was the dependence of the ANN-predictor utilized (precision, recall), on the size of the search space deployed. Additionally, a dichotomy involving regular and potential confounding sequences was observed for the cumulative differences between the profile HMM scores (inter- and intra-sequences). Whilst, the implications and relevance of these findings are significant, much of this work was based on analysis of extant data, *i.e.*, curated and pre-selected sequences; which could preclude usage of this filter as a general

ACCEPTED MANUSCRIPT

screen. Here, I present a mathematical and analytical exposition of the hypothesis, that the scores of these pHMMs ($\Delta pHMMs$) could be utilized to compute an index, which can unambiguously characterize a confounding sequence. Additionally, the numerical description of the bounds and limits, and the utility of this expression in subfamily assignment has been explored.

METHODS

Data preprocessing and primary grouping

The nomenclature for the mathematical descriptor and algorithm are in accordance with earlier work (Kundu and Sharma 2016). Briefly, an indexed superset of the HMMs of the probable biological functions/ subfamilies ($f: \boldsymbol{K} \leftrightarrow \{1,2 \ldots .n\}, K \subset \mathbb{R}$) that an unknown protein sequence is posited to possess is defined $\{k_a \in \boldsymbol{K}, 1 \leq a \leq |\boldsymbol{K}|\}(Def.1)$. Here, each k_a describes the HMM-score (highest/ lowest E-value) of a multiple sequence alignment (MSA) of a set of template sequences that exhibits or is likely to manifest that particular function. The unknown sequence is then queried with the combined set of HMMs (profile HMMs, pHMM), the raw scores of which are then grouped and modified. Finally, a numerical value is assigned to the pairs-of-pairs $\{\alpha_n(k_i k_j k_l k_m) \in \boldsymbol{L}: 1 \leq n \leq |\boldsymbol{L}|; k_i, k_j, k_l, k_m \in \boldsymbol{K}\}$, which is then utilized to determine the grade of the protein sequence:

$$|\boldsymbol{K}| = 3; \boldsymbol{K} = \{k_1, k_2, k_3\} \qquad \text{Def. 1}$$

$$|\boldsymbol{M}| = C(|\boldsymbol{K}|, 2) = \binom{|\boldsymbol{K}|}{2} = 3; \boldsymbol{M} = \{(k_1, k_2), (k_2, k_3), (k_1, k_3)\} \qquad \text{Def. 2}$$

$$\left|\boldsymbol{L}_{|\boldsymbol{K}|}\right| = C(|\boldsymbol{M}|,2) = \binom{|\boldsymbol{M}|}{2} = C\left(\binom{|\boldsymbol{K}|}{2}, 2\right) = \binom{\binom{|\boldsymbol{K}|}{2}}{2} = 3$$

$$\boldsymbol{L} = \{\alpha_1((k_1, k_2), (k_2, k_3)), \alpha_2((k_2, k_3), (k_1, k_3)), \alpha_3((k_1, k_2), (k_1, k_3))\} \qquad \text{Def. 3}$$

$\boldsymbol{K}$ $:=$ $Profile\ HMM\ scores\ of\ a\ protein\ sequence$
$\boldsymbol{M}$ $:=$ $Pairs\ of\ profile\ HMM\ scores\ of\ a\ protein\ sequence$
$\boldsymbol{L}$ $:=$ $Pairs - of - pairs\ (POP)\ of\ profile\ HMM\ scores\ of\ a\ protein\ sequence$
α $:=$ $Computed\ statistical\ descriptor\ of\ POP\ profile\ HMM\ scores\ for\ a\ protein\ sequence$

Secondary clustering of $\boldsymbol{L}_{|\boldsymbol{K}|}$

The processed HMM scores of the superset ($\boldsymbol{L}_{|\boldsymbol{K}|}$) are partitioned into subsets $\boldsymbol{F}$ and $\boldsymbol{G}$ ($\boldsymbol{F},\boldsymbol{G} \subset \boldsymbol{L}, |\boldsymbol{F}| + |\boldsymbol{G}| = |\boldsymbol{L}|$)**.** This formulation is based on the treating the POPs as combinations of individual elements. Thus, for each subset of POPs ($\alpha_{\boldsymbol{n}} \in \boldsymbol{F},\boldsymbol{G}$), the elemental composition, *i.e.*, the probable set of biological functions ($\boldsymbol{K}$) can either be arranged four at a time with zero redundancy $\{\alpha_n(k_i,k_j,k_l,k_m) \in \boldsymbol{F}: 1 \le n \le |\boldsymbol{F}|, k_i \ne k_j \ne k_l \ne k_m; k_i,k_j,k_l,k_m \in \boldsymbol{K}\}(\text{Def. }4)$, or three at a time with a single redundancy $\{\alpha_n(k_i,k_j,k_l,k_m) \in \boldsymbol{G}: 1 \le n \le |\boldsymbol{G}|, k_i \ne k_j \ne k_l \;\&\; k_m = \{k_i,k_j,k_l\}; k_i,k_j,k_l,k_m \in \boldsymbol{K}\ \}(\text{Def. }5)$. The formulae to compute the cardinality of $\boldsymbol{F}$ and $\boldsymbol{G}$ is:

$$|\boldsymbol{F}| = (3)\binom{|\boldsymbol{K}|}{4} \qquad \text{Eq. 1}$$

$$|\boldsymbol{G}| = (3)\binom{|\boldsymbol{K}|}{3} \qquad \text{Eq. 2}$$

Clearly, for $|\boldsymbol{K}| = 3$, $|\boldsymbol{L}| = 3\ (Def.3)$. Since, $\boldsymbol{F},\boldsymbol{G} \subset \boldsymbol{L}$, and $|\boldsymbol{G}| = 3\ (Eq.2)$, it follows that $|\boldsymbol{F}| = 0\ (|\boldsymbol{F}| + |\boldsymbol{G}| = |\boldsymbol{L}|)$. Similarly, for $|\boldsymbol{K}| \ge 4$, $|\boldsymbol{F}| \ge 3$ and $|\boldsymbol{G}| \ge 12$ $(\text{Eqs. }1,2)$.

Tertiary subgroups of $\boldsymbol{L}_{|\boldsymbol{K}|}$ and numerical transformation of POPs

The POP-scores in **G** ($\alpha_n(\boldsymbol{G})$), as defined *vide supra* are then rearranged $\{\alpha_{ab}(\boldsymbol{G}) \in \boldsymbol{G}^*: 1 \le a \le |\boldsymbol{K}|, 1 \le b \le \lambda\}(\text{Def. }6)$. This bijection ($f: \alpha_n(\boldsymbol{G}) \leftrightarrow \alpha_{ab}(\boldsymbol{G})$), partitions $\boldsymbol{G}^*$ into $|\boldsymbol{K}|$-clusters of vectors ($Dim = \lambda = {}^{|\boldsymbol{G}^*|}/_{|\boldsymbol{K}|}$)$(\text{Eq. }3)$, wherein each cluster/subset represents a grouping of self-referencing POPs spanning a specific biological function ($\alpha_{ab}(\boldsymbol{G})$). Each POP is then scored individually using the statistical descriptors outlined *vide infra* $(\text{Eqs. }4, 14{,}15)$.

$$f(\alpha_{ab}) = \begin{cases} {}^{\lfloor \alpha_{ab} \rfloor}/_{\alpha_{ab}}, & 0 \le \alpha_{ab} < 1 \\ {}^{\alpha_{ab}}/_{\lfloor \alpha_{ab} \rfloor}, & \alpha_{ab} \ge 1 \end{cases} \qquad \text{Eq. 4a}$$

$$f(\alpha_{ab}) = \begin{cases} 0, & 0 \le \alpha_{ab} < 1 \\ 1, & \alpha_{ab} \ge 1 \end{cases} \qquad \text{Eq. 4b}$$

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

$\alpha \quad := \quad Computed\ statistical\ descriptor\ of\ \ POP\ of\ profile\ HMM\ scores\ for\ a\ protein\ sequence$
$a \quad := \quad Subfamily\ index\ of\ a\ particular\ POP = 1 \leq a \leq |\boldsymbol{K}|$
$b \quad := \quad Vector\ component\ of\ a\ particular\ POP = 1 \leq b \leq \lambda$

The subfamily specific cluster of POPs $\{\alpha_{ab}(\boldsymbol{G}) = \bigcup_{a=1}^{a=|\boldsymbol{K}|}(\bigcup_{1\leq b\leq\lambda}(\alpha_{ab}) \in \boldsymbol{G}^*: 1 \leq a \leq |\boldsymbol{K}|, 1 \leq b \leq \lambda\}$ (Def. 7) are thus transformed into equivalent subsets of binary vectors $\{0,1\}$ of dimension (λ). The superset thus formed is denoted by

$\{f(\boldsymbol{G}^*) = f\big(\alpha_{ab}(\boldsymbol{G})\big) = f(\bigcup_{a=1}^{a=|\boldsymbol{K}|}(\bigcup_{1\leq b\leq\lambda}(\alpha_{ab})) = \bigcup_{a=1}^{a=|\boldsymbol{K}|}(\bigcup_{1\leq b\leq\lambda} f(\alpha_{ab})) \in \boldsymbol{H}:\}$ (Def. 8).

Computing the index of sequence suitability

The notation for the computed component vectors of each probable biological function is simplified $(\phi_{k_a} \equiv f\big(\alpha_{ab}(\boldsymbol{G})\big) \in \boldsymbol{H}: 1 \leq a \leq |\boldsymbol{K}|, 1 \leq b \leq \lambda, k_a \in \boldsymbol{K})$(Def. 9) for a sequence. Additionally, a theoretical vector space $(\boldsymbol{V} \subset \mathbb{R}^{\boldsymbol{\lambda}})$ of combinatorially arranged clusters of λ-binary vectors that accounts for every possible outcome of a specific biological function $(\phi_j \in \boldsymbol{V}: 1 \leq j \leq \lambda)$ of dimension $\mathbf{2}^{|G^*|/|\boldsymbol{K}|} = 2^{\lambda}$ is defined (Def. 10; Eq. 5). These are combined as in $Eq.$ 6, and the result is incorporated into the formula (Eq. 7) and used to assess the state of an unknown protein sequence:

$$\gamma_{k_a} = \max_{1\leq j\leq 2^{\lambda}}\left(\phi_{k_a}(\phi_j)^T\big/_{\lambda}\right) = \begin{cases} 0, & 0 \leq \gamma_{k_a} < 1 \\ 1, & \gamma_{k_a} = 1 \end{cases} \qquad \text{Eq. 6}$$

$$\theta_{seq} = {\textstyle\sum} \gamma_{k_a}\big/_{|K|} \qquad \text{Eq. 7}$$

$\gamma_{k_a} \quad := \quad Score\ of\ unique\ subfamilies/profiles$
$\phi_{k_a} \quad := \quad Computed\ vector\ \ of\ {k_a}^{th}\ biological\ function$
$\phi_j \quad := \quad Component\ vector\ of\ theoretical\ vector\ space$
$\lambda \quad = \quad {|G^*|}\big/_{|\boldsymbol{K}|}$
$|\boldsymbol{K}| = N_{subfam} \quad := \quad Total\ number\ of\ predicted\ subfamilies/profiles$
$\theta_{seq} \quad := \quad Suitability\ index\ of\ protein\ sequence$

Algorithm to compute sequence suitability index

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Step 1: Define the superset of probable biological functions that an unknown protein sequence needs to be evaluated ($\boldsymbol{K}$) against, and complete the primary, secondary and tertiary clustering. This defines the sets **L**, **F**, and **G.**

Step 2: Ignore the subsets $\boldsymbol{F} \subset \boldsymbol{L}_{|\boldsymbol{K}|}$, *i.e.,* $\boldsymbol{L}_{|\boldsymbol{K}|} - \boldsymbol{F}$. Here, $|\boldsymbol{F}| = \left|\boldsymbol{L}_{|\boldsymbol{K}|-\boldsymbol{1}}\right|, |\boldsymbol{K}| \geq 4$ Eq. 8

Step 3: Include the following subsets $\boldsymbol{G} \subset \boldsymbol{L}_{|\boldsymbol{K}|}$. Here, $|\boldsymbol{G}| = \left(\left|\boldsymbol{L}_{|\boldsymbol{K}|}\right| - \left|\boldsymbol{L}_{|\boldsymbol{K}|-\boldsymbol{1}}\right|\right), |\boldsymbol{K}| \geq 4$ Eq. 9

Step 4: Numerically transform the rearranged tertiary collection of POPs of the subset $\boldsymbol{G}^{*}$**,** *i.e.*, $\alpha_{ab}(\boldsymbol{G})$ into a biological function specific combinatorial collection of computationally viable vectors $f(\alpha_{ab}(\boldsymbol{G}))$ in accordance with $(\text{Defs. } 6-8; \text{Eqs. } 4{,}14{,}15)$.

Step 5: Assess the contribution of each biological function relevant vector (γ_{k_a}) individually and combine the result $(\text{Defs. } 9{,}10; \text{Eqs. } 5{,}6)$.

Step 6: Calculate the suitability index of a specific protein sequence with $(Eq. 7)$. Thus, a suitable sequence ($\theta_{seq} = 1.00; \sum \gamma_{k_a} = N_{subfam}$; Eq. 10a), while a confounding sequence is indicated by ($0.00 \leq \theta_{seq} < 1.00; \sum \gamma_{k_a} < N_{subfam}$; Eq. 10b).

Sequence specific probabilities of subfamily prediction

This was computed using the binomial distribution of an experiment which consisted of all possible trials for a particular subfamily/ profile of the sequence of interest, or as defined previously the vector space ($\boldsymbol{V}$) of every probable combination of $\{0{,}1\}$ that could theoretically characterize a subfamily of a protein sequence $(\text{Def. } 10; \text{Eqs. } 3{,}5)$. Here, each trial ($T$) is modeled as a composite of λ-binary outcomes. This was formulated as:

$\mathcal{B}_{k_a}(ST; T, pST); 1 \leq a \leq |\boldsymbol{K}|, k_a \in \boldsymbol{K}, |\boldsymbol{K}| \geq 3$ Eq. 11

$$\begin{array}{ccl} k_a & := & Specific\ biological\ function\ /Subfamily \\ |\boldsymbol{K}| & := & Total\ number\ of\ subfamilies \\ ST & := & Successful\ trial\ \equiv [1_1 1_2 .. 1_\lambda] \\ T & := & Trials\ for\ a\ subfamily = 2^\lambda \\ pST & := & Probability\ of\ a\ successful\ trial\ (^{ST}/_{T}) \end{array}$$

RESULTS AND DISCUSSION

Modelling the probable functional space of proteins

The putative function(s) of a protein can be modeled as the joint probability of its constituent predicted functions (subfamilies/ profiles). Consider the arbitrary four classes of enzymatic activity a protein sequence is purported to possess, *i.e.*,

sequence := (profile class k_1)(profile class k_2)(profile class k_3)(profile class k_4). This may be represented as:

$$\boldsymbol{Activity}\ (\boldsymbol{sequence}\ ) = \prod_{a=1}^{a=|\boldsymbol{K}|} (\delta_{k_a})(k_a) \qquad \text{Eq. 12}$$

$$\begin{array}{ccl} \delta & := & Normalized\ values\ of\ averaged\ raw\ profile\ HMM\ scores\ of\ sequence \\ k_a & := & Specific\ biological\ function\ /Subfamily \\ N_{subfam} = |\boldsymbol{K}| & = & Total\ number\ of\ subfamilies \end{array}$$

Case 1: $sequence\ (x) := (0.12)(k_1)\ (0.05)(k_2)\ (0.30)(k_3)(0.575)(k_4)$

Case 2: $sequence\ (y) := (0.03)(k_1)\ (0.90)(k_2)\ (0.04)(k_3)(0.33)(k_4)$

Case 3: $sequence\ (z) := (0.02)(k_1)\ (0.03)(k_2)\ (0.85)(k_3)(0.145)(k_4)$

Case 4: $sequence\ (x') := (0.45)(k_1)\ (0.45)(k_2)\ (0.05)(k_3)(0.95)(k_4)$

Case 5: $sequence\ (y') := (0.25)(k_1)(0.30)(k_2)\ (0.40)(k_3)(0.05)(k_4)$

Case 6: $sequence\ (z') := (0.24)(k_1)\ (0.26)(k_2)\ (0.25)(k_3)(0.25)(k_4)$

The expression (Eq. 12) is of limited utility. The combined value does not indicate anything about the individual functions (subfamilies), neither does it offers any insight into the contribution of each predicted function. Further, sequences with profiles with one or more closely spaced HMM scores (Cases 2- 6), too

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

cannot be distinguished. An attempt to resolve this resulted in an expression involving the rearrangement of the HMM scores and their subsequent scoring using a modification of the Z-score (Kundu and Sharma 2016). This sequence specific β-value (Eq. 13) was used in tandem with available empirical data to numerically determine a threshold, which was then used to screen sequences. Although, the approach improved the accuracy of predictions, the effect was indirect and was due to ignoring sequences that were below a certain threshold (Kundu and Sharma 2016). Rewriting the β-value for a protein sequence with the current notation:

$$\beta = \left({}^{1}\!/_{2}\right) \sum_{a=1}^{a=|\boldsymbol{K}|} \sum_{b=1}^{b=\lambda} \alpha_{ab}\left(k_i, k_j, k_l, k_m\right) = \left({}^{1}\!/_{2}\right)\left(\sum_{a=1}^{a=|\boldsymbol{K}|} \sum_{b=1}^{b=\lambda} \alpha_{ab}\ \left(k_i, k_j, k_l, k_m\right)\right); k_i, k_j, k_l, k_m \in \boldsymbol{K} \quad \text{Eq. 13}$$

Inadequacy of the $\boldsymbol{\beta}$-value in determining sequence suitability

An analytical treatment of $Eq. 4$ highlights the aforementioned limitations in the absence of corroborating data.

$$\alpha_{ab}(k_i, k_j, k_l, k_m) = \left(\left(|\mu_{k_ik_j} - \mu_{k_lk_m}|\right)\Big/_{\omega}\right)\left(Z^{k_ik_j,k_lk_m}\right), 1 \le a \le |\boldsymbol{K}|, 1 \le b \le \lambda \qquad \text{Eq. 14}$$

$$\begin{array}{lcl}
k_i, k_j, k_l, k_m & \in & \boldsymbol{K} \\
\mu & := & average\ value\ of\ a\ POP\ component \\
\omega & := & scaling\ factor = 100 \\
\mathrm{Z} & := & Z\ score\ for\ each\ POP = \left(\left(|\mu_{k_ik_j} - \mu_{k_lk_m}|\right) \Big/ \left(\sqrt[2]{\left(\sigma^2_{k_ik_j} + \sigma^2_{k_lk_m}\right)\Big/_{\tau^{k_ik_j,k_lk_m}}} \right) \right) \\
\sigma^2 & := & variance\ for\ a\ POP\ component \\
\tau & := & component\ members\ of\ each\ POP = 2
\end{array}$$

Assume $\alpha_{ab} \in [1.00, \infty)$. Then Eq. 14 may be re-written as

ACCEPTED MANUSCRIPT

$$\begin{aligned}
\left(|\mu_{k_ik_j} - \mu_{k_lk_m}|\right)^2 &= (\alpha_{ab})(\omega)\left(\sqrt[2]{\left(\sigma^2_{k_ik_j} + \sigma^2_{k_lk_m}\right)\Big/_{\tau^{k_ik_j,k_l,k_m}}}\right) \\
&= (\alpha_{ab})(\omega)(\sqrt{\overline{\sigma^2_{k_lk_j,k_l,k_m}}}) \qquad \text{Eq. 15} \\
&= (\alpha_{ab})(\omega)(\overline{\sigma_{k_lk_j,k_l,k_m}}) \\
|\mu_{k_ik_j} - \mu_{k_lk_m}| &= \sqrt{(\alpha_{ab})(\omega)(\overline{\sigma_{k_lk_j,k_l,k_m}})}
\end{aligned}$$

Clearly, as $\alpha_{ab} \to \infty, |\mu_{k_ik_j} - \mu_{k_lk_m}| \to \infty$

Similarly, for $\alpha_{ab} \in (0.00,1.00)$ and as $\alpha_{ab} \to 0.00, |\mu_{k_ik_j} - \mu_{k_lk_m}| \to 0.00$

Using the above,

$\beta \geq \binom{|\boldsymbol{K}|}{2}, \alpha_{ab} \in [1.00, \infty)$ Eq. 16a

$\beta < \binom{|\boldsymbol{K}|}{2}, \alpha_{ab} \in (0.00, 1.00)$ Eq. 16b

Despite, the definition of bounds $(\text{Eqs.}\, 14 - 16)$ for the β-value, there is minimal information on the contributory influence of each pHMM on the overall score, an unresolved problem as in the case of $Eq. 12$. Additionally, since β-value is an expression of summation $(Eq. 13)$, it is likely, that one or more pHMM scores are not well spaced (Cases 2-6). This would imply that even if the β-value for a sequence (β_{seq}) exceeded the set threshold $(\text{Eq.}\, 15\text{a})$, there is no information on the individual POP scores (α_{ab}) which could $\alpha_{ab} \to 0.00$ (Cases 2-6)$(\text{Eq.}\, 15\text{b})$. Whilst, the presence of sequences with available kinetic data could certainly resolve these, this would limit usage of the filter to specific instances of proteins sequences (Kundu and Sharma 2016).

Mapping the $\boldsymbol{\Delta pHMMs}$ to a favorable profile

The formulae, expressions, and algorithm $(\text{Defs.}\, 1 - 10, \text{Eqs.}\, 1 - 15; \text{steps}\ 1 - 6)$, and depends on the partitioning of the POPs $(\boldsymbol{M}, \boldsymbol{L}, \boldsymbol{F}, \boldsymbol{G})$. The following worked example for Case 1, *i.e.*, sequence *(x)* illustrates this:

Set of functions $(K = \{k_1, k_2, k_3, k_4\}$ with cardinality $|K| = 4; \text{Def.}\, 1)$

Set of pair-of-pairs (L with cardinality $|L_4| = 15; \text{Defs.}\, 1 - 3$)

Consider the following relevant subsets$|L_4| - |L_3| = 12; \text{Defs.}\, 4 - 8$

$$\alpha_{11}(\boldsymbol{G}) = \{(k_1, k_3), (k_2, k_4): k_1, k_2, k_3, k_4 \in \mathbb{R}\} = \{(k_1, k_3), (k_1, k_4): k_1 = k_2, k_1 \neq k_3 \neq k_4 \text{ and } k_1, k_3, k_4 \in \mathbb{R}\}$$

$$\alpha_{21}(\boldsymbol{G}) = \{(k_1, k_3), (k_2, k_4): k_1, k_2, k_3, k_4 \in \mathbb{R}\} = \{(k_2, k_3), (k_2, k_4): k_2 = k_1, k_2 \neq k_3 \neq k_4 \text{ and } k_2, k_3, k_4 \in \mathbb{R}\}$$

$$\alpha_{12}(\boldsymbol{G}) = \{(k_1, k_2), (k_3, k_4): k_1, k_2, k_3, k_4 \in \mathbb{R}\} = \{(k_1, k_2), (k_1, k_4): k_1 = k_3, k_1 \neq k_2 \neq k_4 \text{ and } k_1, k_2, k_4 \in \mathbb{R}\}$$

$$\alpha_{31}(\boldsymbol{G}) = \{(k_1, k_2), (k_3, k_4): k_1, k_2, k_3, k_4 \in \mathbb{R}\} = \{(k_3, k_2), (k_3, k_4): k_3 = k_1, k_3 \neq k_2 \neq k_4 \text{ and } k_2, k_3, k_4 \in \mathbb{R}\}$$

$$\alpha_{13}(\boldsymbol{G}) = \{(k_1, k_2), (k_4, k_3): k_1, k_2, k_3, k_4 \in \mathbb{R}\} = \{(k_1, k_2), (k_1, k_3): k_1 = k_4, k_1 \neq k_2 \neq k_3 \text{ and } k_1, k_2, k_3 \in \mathbb{R}\}$$

$$\alpha_{41}(\boldsymbol{G}) = \{(k_1, k_2), (k_4, k_3): k_1, k_2, k_3, k_4 \in \mathbb{R}\} = \{(k_4, k_2), (k_4, k_3): k_4 = k_1, k_4 \neq k_2 \neq k_3 \text{ and } k_2, k_3, k_4 \in \mathbb{R}\}$$

$$\alpha_{22}(\boldsymbol{G}) = \{(k_2, k_1), (k_3, k_4): k_1, k_2, k_3, k_4 \in \mathbb{R}\} = \{(k_2, k_1), (k_2, k_4): k_2 = k_3, k_2 \neq k_1 \neq k_4 \text{ and } k_1, k_2, k_4 \in \mathbb{R}\}$$

$$\alpha_{32}(\boldsymbol{G}) = \{(k_2, k_1), (k_3, k_4): k_1, k_2, k_3, k_4 \in \mathbb{R}\} = \{(k_3, k_1), (k_3, k_4): k_3 = k_2, k_3 \neq k_1 \neq k_4 \text{ and } k_1, k_3, k_4 \in \mathbb{R}\}$$

$$\alpha_{23}(\boldsymbol{G}) = \{(k_2, k_1), (k_4, k_3): k_1, k_2, k_3, k_4 \in \mathbb{R}\} = \{(k_2, k_1), (k_2, k_3): k_2 = k_4, k_2 \neq k_1 \neq k_3 \text{ and } k_1, k_2, k_3 \in \mathbb{R}\}$$

$$\alpha_{42}(\boldsymbol{G}) = \{(k_2, k_1), (k_4, k_3): k_1, k_2, k_3, k_4 \in \mathbb{R}\} = \{(k_4, k_1), (k_4, k_3): k_4 = k_2, k_4 \neq k_1 \neq k_3 \text{ and } k_1, k_3, k_4 \in \mathbb{R}\}$$

$$\alpha_{33}(\boldsymbol{G}) = \{(k_4, k_1), (k_3, k_2): k_1, k_2, k_3, k_4 \in \mathbb{R}\} = \{(k_3, k_1), (k_3, k_2): k_3 = k_4, k_3 \neq k_1 \neq k_2 \text{ and } k_1, k_2, k_3 \in \mathbb{R}\}$$

$$\alpha_{43}(\boldsymbol{G}) = \{(k_4, k_1), (k_3, k_2): k_1, k_2, k_3, k_4 \in \mathbb{R}\} = \{(k_4, k_1), (k_4, k_2): k_4 = k_3, k_4 \neq k_1 \neq k_2 \text{ and } k_1, k_2, k_4 \in \mathbb{R}\}$$

The cumulative profiles $\phi_{k_a}(\boldsymbol{G}): 1 \leq a \leq 4,\ \boldsymbol{V} = \{[111], [100], [110], [101], [001], [011], [010], [000]\}$

$\phi_{k_1}(\boldsymbol{G}) = [\alpha_{11}(\boldsymbol{G})\ \alpha_{12}(\boldsymbol{G})\ \alpha_{13}(\boldsymbol{G})], f\big(\alpha_1(\boldsymbol{G})\big) = [f(\alpha_{11}(\boldsymbol{G})\ f(\alpha_{12}(\boldsymbol{G})\ f(\alpha_{13}(\boldsymbol{G}))] = [111]\ \in \boldsymbol{V}$ (Defs. 9, 10)

$\phi_{k_2}(\boldsymbol{G}) = [\alpha_{21}(\boldsymbol{G})\ \alpha_{22}(\boldsymbol{G})\ \alpha_{23}(\boldsymbol{G})], f\big(\alpha_2(\boldsymbol{G})\big) = [f(\alpha_{21}(\boldsymbol{G})\ f(\alpha_{22}(\boldsymbol{G})\ f(\alpha_{23}(\boldsymbol{G}))] = [111]\ \in \boldsymbol{V}$(Defs. 9, 10)

$\phi_{k_3}(\boldsymbol{G}) = [\alpha_{31}(\boldsymbol{G})\ \alpha_{32}(\boldsymbol{G})\ \alpha_{33}(\boldsymbol{G})], f\big(\alpha_3(\boldsymbol{G})\big) = [f(\alpha_{31}(\boldsymbol{G})\ f(\alpha_{32}(\boldsymbol{G})\ f(\alpha_{33}(\boldsymbol{G}))] = [111]\ \in \boldsymbol{V}$(Defs. 9, 10)

$\phi_{k_4}(\boldsymbol{G}) = [\alpha_{41}(\boldsymbol{G})\ \alpha_{42}(\boldsymbol{G})\ \alpha_{43}(\boldsymbol{G})], f\big(\alpha_4(\boldsymbol{G})\big) = [f(\alpha_{41}(\boldsymbol{G})\ f(\alpha_{42}(\boldsymbol{G})\ f(\alpha_{43}(\boldsymbol{G}))] = [111]\ \in \boldsymbol{V}$(Defs. 9, 10)

Since, $\gamma_{k_1} = \gamma_{k_2} = \gamma_{k_3} = \gamma_{k_4} = 1.00$ Eq. 6

$\theta_{seq}(x) = {}^{\sum \gamma_{k_a}}/_{4} = 1.00$ Eq. 7

$\mathcal{B}_{k_1}(x)(1; 8{,}0.125) = \mathcal{B}_{k_2}(x)(1; 8{,}0.125) = \mathcal{B}_{k_3}(x)(1; 8{,}0.125) = \mathcal{B}_{k_4}(x)(1; 8{,}0.125) = 0.3927$ Eq. 11

Similarly, for the hypothetical cases 2-6:

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

$\theta_{seq}(y), \theta_{seq}(z), \theta_{seq}(x'), \theta_{seq}(y') < 1.00$
$\theta_{seq}(z') = 0.00$

Thus, using the above, an arbitrary protein sequence (x) (Case 1), could be included with confidence into a defined training set, as opposed to sequences (y, z, x', y', z') (Cases 2-6) (Fig. 1a).

Significance of sequence filtering in subfamily classification

The probability whether a trial of subfamilies/ profiles is deemed successful imposes a stringency on the outcome of an experiment involving the association of a specific profile/ subfamily with all the others. Since, there is at most only one such trial for a particular profile, the computed probability is consistent $(\mathcal{B}_{k_a} \in [0.367879, 0.50])$ for values $(3 \leq |\boldsymbol{K}| \leq 9)$ (Table 1, Fig. 1b). A direct comparison of these also suggests that the data can be approximated with a polynomial function:

$$y_{poly} = 3E-05x^6 - 0.0014x^5 + 0.0238x^4 - 0.2213x^3 + 1.1581x^2 - 3.232x + 4.1302, R^2 = 1.00 \text{ Eq. 16}$$

$y \quad := \quad Binomial\ probabilty\ of\ a\ predicted\ subfamily\ (k_a \in \boldsymbol{K})\ of\ a\ sequence = \mathcal{B}_{k_a}$
$x \quad := \quad Total\ number\ of\ predicted\ subfamilies = |\boldsymbol{K}|$

The numerical characterization and identification of confounding sequences has important consequences for subfamily classification, and indirectly, phylogenetic analysis. In terms of sequence identity, a confounding sequence may also be defined as, one with an identity most proximal to the lowest alignment score in a pairwise comparison of subfamily members $(x' \in X; y' \in Y; z' \in Z)$ (Fig. 1a). This, would imply the existence of a set whose members are localized at the boundary of their individual clusters, and can therefore be considered interchangeably with either of the parent clusters. Mathematically, this can be elaborated as:

Assume $x', x_1, x_2 \in X, y', y_1, y_2 \in Y, z', z_1, z_2 \in Z$ (Fig. 1a)

Then, $x', y', z' \in XYZ$ (Definition) (Fig. 1a)

If $g(t)$ represents the score of a global alignment of sequence t, then $g(x') \approx g(y') \approx g(z')$(Definition), it follows that,

$\{g(x'), g(x_1), g(x_2)\} = \{g(y'), g(z'), g(x_1), g(x_2)\} \in g(X)$

ACCEPTED MANUSCRIPT

$\{g(y'), g(y_1), g(y_2)\} = \{g(x'), g(z'), g(y_1), g(y_2)\} \in g(Y)$

$\{g(z'), g(z_1), g(z_2)\} = \{g(x'), g(y'), g(z_1), g(z_2)\} \in g(Z)$

Clearly, a classification based on the above is likely to attenuate the accuracy of any prediction schema. The use of pHMMs, in contrast, provides a scaffold to directly compare profiles across protein sequences. Since, HMM scores are dependent on the similarity of local sequence segments across the sequence, this approach would automatically entail a region wise comparison across a sequence of interest. This may be written as:

If $\mathrm{h}(t)$ represents the set of HMM scores of a sequence t,

Case 7: If, $h(x') \neq h(y') \neq h(z')$, then,

$\{h(x'), h(x_1), h(x_2)\} \in h(X)$

$\{h(y'), h(y_1), h(y_2)\} \in h(Y)$

$\{h(z'), h(z_1), h(z_2)\} \in h(Z)$

Case 8: If however, $h(x') \approx h(y') \approx h(z')$, then

$\{h(x'), h(x_1), h(x_2)\} = \{h(y'), h(z'), h(x_1), h(x_2)\} \in h(X)$

$\{h(y'), h(y_1), h(y_2)\} = \{h(x'), h(z'), h(y_1), h(y_2)\} \in h(Y)$

$\{h(z'), h(z_1), h(z_2)\} = \{h(x'), h(y'), h(z_1), h(z_2)\} \in h(Z)$

Thus, for a confounding sequence, *i.e.*, with one or more closely spaced pHMM scores the specificity degenerates and results in the nonspecific association with pre-defined clusters (Case 8) ($h(X) \approx g(X); h(Y) \approx g(Y); h(Z) \approx g(Z)$). The pHMM scores of a sequence are clearly superior to simple global alignment similarity scores of the same, since they allow a bifurcation on the basis of regional similarity of a sequence. In other words, sequences with matching global similarity scores are still distinguishable when a comparison is made of their functionally relevant specific sequence segments. Additionally, the usage of established statistical methods for the necessary computations, ensures that that approach is rigorous. The single most relevant limitation, of this method is the subtraction of suboptimal sequences

ACCEPTED MANUSCRIPT

from the search space. This can lead to a significant reduction in sample size, and, parallels the observations made earlier (Kundu and Sharma 2016).

CONCLUDING REMARKS

The availability of a 3D structure in tandem with kinetic and mutagenesis data is supportive of function, and is therefore, invaluable in any annotation schema. Nevertheless, given the large volume of publically available uncurated and uncharacterized data repositories, there is a need for analytic measures that can not only improve our comprehension of the underlying system, while at the same time provide researchers with reliable tools to initiate confirmatory laboratory work. The sequence specific index presented here is independent of empirical data and can characterize a confounding sequence with a high degree of confidence. Additionally, it can be utilized as a generic filter to define high quality templates of sequences, and thereby, assist investigators infer molecular ancestry, and indirectly, function.

COMPETING INTERESTS

The author declares no competing financial interests.

AUTHORS' CONTRIBUTION

SK formulated, developed and evaluated the formulae and filters, carried out the mathematical and computational analysis, wrote all necessary code, and the manuscript.

FUNDING

The author(s) received no funding for this work.

References

ACCEPTED MANUSCRIPT

Chung D, Young J, Cha M, Brunecky R, Bomble YJ, Himmel ME, Westpheling J (2015) Expression of the Acidothermus cellulolyticus E1 endoglucanase in Caldicellulosiruptor bescii enhances its ability to deconstruct crystalline cellulose Biotechnology for biofuels 8:113 doi:10.1186/s13068-015-0296-x

Fukumori F, Hausinger RP (1993) Alcaligenes eutrophus JMP134 "2,4-dichlorophenoxyacetate monooxygenase" is an alpha-ketoglutarate-dependent dioxygenase Journal of bacteriology 175:2083-2086

Hogan DA, Auchtung TA, Hausinger RP (1999) Cloning and characterization of a sulfonate/alpha-ketoglutarate dioxygenase from Saccharomyces cerevisiae Journal of bacteriology 181:5876-5879

Jones DT (1999a) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences Journal of molecular biology 287:797-815 doi:10.1006/jmbi.1999.2583

Jones DT (1999b) Protein secondary structure prediction based on position-specific scoring matrices Journal of molecular biology 292:195-202 doi:10.1006/jmbi.1999.3091

Khater S, Mohanty D (2015) In silico identification of AMPylating enzymes and study of their divergent evolution Scientific reports 5:10804 doi:10.1038/srep10804

Kundu S (2012) Distribution and prediction of catalytic domains in 2-oxoglutarate dependent dioxygenases BMC research notes 5:410 doi:10.1186/1756-0500-5-410

Kundu S (2015) Unity in diversity, a systems approach to regulating plant cell physiology by 2-oxoglutarate-dependent dioxygenases Frontiers in plant science 6:98 doi:10.3389/fpls.2015.00098

Kundu S, Sharma R (2016) In silico Identification and Taxonomic Distribution of Plant Class C GH9 Endoglucanases Frontiers in plant science 7:1185 doi:10.3389/fpls.2016.01185

Lees J, Yeats C, Redfern O, Clegg A, Orengo C (2010) Gene3D: merging structure and function for a Thousand genomes Nucleic acids research 38:D296-300 doi:10.1093/nar/gkp987

Letunic I et al. (2002) Recent improvements to the SMART domain-based sequence annotation resource Nucleic acids research 30:242-244

Mount DW (2008) Comparison of the PAM and BLOSUM Amino Acid Substitution Matrices CSH protocols 2008:pdb ip59 doi:10.1101/pdb.ip59

Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains Nucleic acids research 26:320-322

Urbanowicz BR, Catala C, Irwin D, Wilson DB, Ripoll DR, Rose JK (2007) A tomato endo-beta-1,4-glucanase, SlCel9C1, represents a distinct subclass with a new family of carbohydrate binding modules (CBM49) The Journal of biological chemistry 282:12066-12074 doi:10.1074/jbc.M607925200

Table 1: Variation of the Binomial probability of a successful trial for a subfamily

|K|	λ	T	ST	pST	bST
3	1	2	1	0.5	0.5
4	3	8	1	0.125	0.3927
5	6	64	1	0.0156	0.3708
6	10	1024	1	0.001	0.3681
7	15	32768	1	3E-05	0.3679
8	21	2097152	1	5E-07	0.3679
9	28	268435456	1	4E-09	0.3679
10	36	6.872E+10	1	1E-11	UNDEF
11	45	3.518E+13	1	3E-14	UNDEF
12	55	3.603E+16	1	3E-17	UNDEF
13	66	7.379E+19	1	1E-20	UNDEF

ACCEPTED MANUSCRIPT

14	78	3.022E+23	1	3E-24	UNDEF
15	91	2.476E+27	1	4E-28	UNDEF
16	105	4.056E+31	1	2E-32	UNDEF
17	120	1.329E+36	1	8E-37	UNDEF
18	136	8.711E+40	1	1E-41	UNDEF
19	153	1.142E+46	1	9E-47	UNDEF
20	171	2.993E+51	1	3E-52	UNDEF
21	190	1.569E+57	1	6E-58	UNDEF
22	210	1.646E+63	1	6E-64	UNDEF
23	231	3.451E+69	1	3E-70	UNDEF
24	253	1.447E+76	1	7E-77	UNDEF
25	276	1.214E+83	1	8E-84	UNDEF
26	300	2.037E+90	1	5E-91	UNDEF
27	325	6.835E+97	1	1E-98	UNDEF
28	351	4.59E+105	1	2E-106	UNDEF
29	378	6.16E+113	1	2E-114	UNDEF
30	406	1.65E+122	1	6E-123	UNDEF
31	435	8.87E+130	1	1E-131	UNDEF

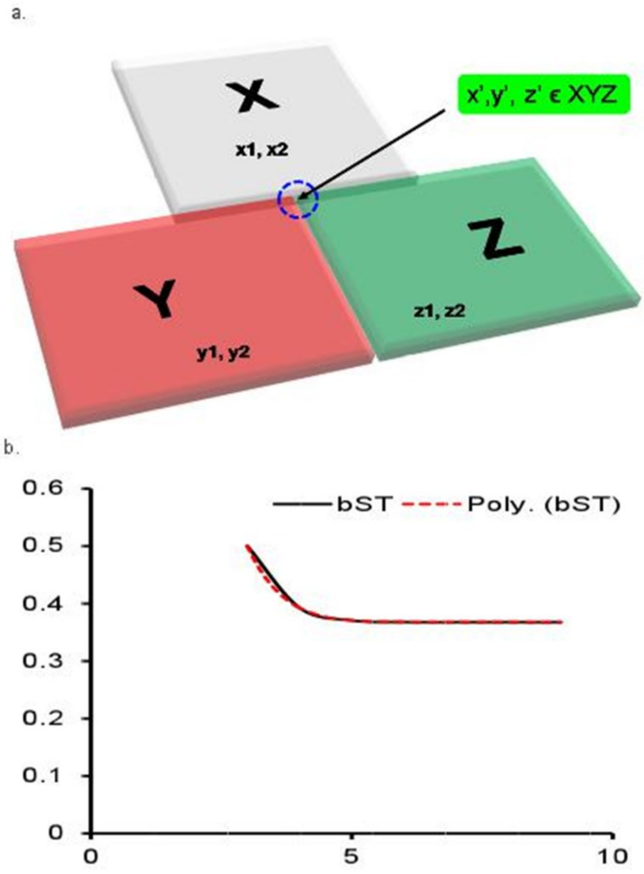
Abbreviations

|**K**|: Total number of subfamilies considered
λ: Dimension of vector
T: Number of predicted trials for a particular subfamily
ST: Number and definition of a successful trial for a subfamily
pST: Probability of a successful trial for a subfamily
bST: Binomial probability of a successful trial for a subfamily ($\mathcal{B}_{k_a}$)

Figure legends

Fig. 1: Impact of confounding sequences on subfamily assignment. a) Schematic diagram highlighting the implied dual membership of confounding protein sequences (x', y', z') of protein subfamilies (X, Y, Z), and b) Scatter plot (black) of the defined set of all subfamilies $(x = |K| \geq 3)$ with the Binomial probability of success for each subfamily $(y = \mathcal{B}_{k_a})$; and the polynomial approximation of the same (red dotted).

ACCEPTED MANUSCRIPT


a.
X
x1, x2
x',y', z' ϵ XYZ
Y
y1, y2
Z
z1, z2
b.
bST
Poly. (bST)
0.6
0.5
0.4
0.3
0.2
0.1
0
0
5
10

ACCEPTED