# Hidden Markov Models Incorporating Fuzzy Measures and Integrals for Protein Sequence Identification and Alignment

Niranjan P. Bidargaddi, Madhu Chetty, and Joarder Kamruzzaman*

*Gippsland School of Computing and Information Technology, Monash University, Churchill, VIC 3842, Australia.*

**Profile hidden Markov models (HMMs) based on classical HMMs have been widely applied for protein sequence identification. The formulation of the forward and backward variables in profile HMMs is made under statistical independence assumption of the probability theory. We propose a fuzzy profile HMM to overcome the limitations of that assumption and to achieve an improved alignment for protein sequences belonging to a given family. The proposed model fuzzifies the forward and backward variables by incorporating Sugeno fuzzy measures and Choquet integrals, thus further extends the generalized HMM. Based on the fuzzified forward and backward variables, we propose a fuzzy Baum-Welch parameter estimation algorithm for profiles. The strong correlations and the sequence preference involved in the protein structures make this fuzzy architecture based model as a suitable candidate for building profiles of a given family, since the fuzzy set can handle uncertainties better than classical methods.**

**Key words: fuzzy profile HMM, sequence alignment, fuzzy measure, fuzzy integral**

## Introduction

Hidden Markov models (HMMs) are probabilistic models that have been applied in various biological problems. For example, profile HMMs (*1–4*) are used for aligning protein sequences of the same family based on homology through Viterbi algorithm. The parameters of profile HMMs are estimated by MAP and Baum-Welch algorithms (*5*). However, classical HMMs have several limitations. First, HMMs do not capture any higher-order correlations of the amino acids in protein sequences. An HMM assumes that the identity of an amino acid at a particular position is independent of the identity of all other positions (*6*). Second, HMMs are also constrained by the statistical independence assumptions during the formulation of the forward and backward variables that are used to compute the matching scores of an unknown sequence to a known family. Due to such assumptions, the joint measure variables (forward and backward) are decomposed as a combination of two measures defined on amino acid emission probabilities and state probabilities. To relax such assumptions and achieve improved performance and flexibility, Mohamed and Gader (*7*) proposed a fuzzy HMM based on fuzzy measures and integrals. Fuzzy measures are an extension of the classical additive measures, obtained by replacing the additive requirement of classical measures with weaker properties of monotonicity, continuity, and semi-continuity (*8*). Integrals are used to aggregate the fuzzy measures by combining the partial support for a hypothesis from the viewpoint of each information source and the importance of various subsets of sources. This model does not require the assumption of decomposing the measures. It also does not require fixing the lengths of the sequences and the availability of large training datasets as required by classical HMMs in order to learn the transition parameters, thus it offers more flexibility and robustness. The fuzzy HMM has been successfully applied in various domains such as speech processing and image processing (*9–11*).

The fuzzy measure concept for protein sequence analysis was first introduced into profile HMMs in our previous studies (*12*, *13*), which showed improved performance over classical profile HMMs. In this paper, we originally propose a fuzzy Viterbi search algorithm based on Choquet integrals and fuzzy measures in order to overcome the limitations of the classical Viterbi search algorithm that has been used traditionally to align a query sequence to a profile model. It in-

***Corresponding author. E-mail: Joarder.Kamruzzaman@infotech.monash.edu.au**

corporates ascending values of the scores of the neighboring states while calculating the scores for a given state, hence providing better alignment and accurate scores. We also propose a fuzzy Baum-Welch algorithm to relax the statistical independence assumption in the classical Baum-Welch algorithm. Evaluation results obtained with protein sequences from globin and kinase databases demonstrate the superiority of the fuzzy profile HMM over classical models measured in terms of matching scores and alignments.

# Model

## Classical profile HMM

Profiles introduced by Gribskov (*14*) are statistical descriptions of the consensus of multiple sequence alignment, which use position-specific scores for amino acids and position-specific penalties for opening and extending an insertion or deletion. Figure 1 shows the Plan 7 architecure of profile HMM used in software HMMER 2. This architecture differs from the original (Plan 9) Krogh/Hausler architecture (*15*) used in earlier version of HMMER by reducing the number of transitions from 9 to 7, without $D \rightarrow I$ and $I \rightarrow D$ transitions. Profile HMMs capture position-specific information such as how conserved each column of the alignment is and which residues are likely to occur in each column. They are capable of modeling gapped alignments including insertions and deletions, which allows modeling of a complete conserved domain (rather than just a small ungapped motif). If a trusted alignment is not yet known, profile HMMs can be trained from unaligned sequences using Baum-Welch expectation maximization. The profile HMM architecture shown in Figure 1 is characterized by the following parameters:

| | |
|---|---|
| $M_k$ | match state k, with 20 emission probabilities |
| $D_k$ | delete state k, non-emitter |
| $I_k$ | insert state k, with 20 emission probabilities |

It consists of a linear set of match (M), insert (I), and delete (D) states. There is one M state per consensus column in the multiple alignments. Each M state carries a vector of 20 probabilities for scoring the 20 amino acids. Each M state is associated with an I and a D state. The group of three states (M/D/I) at the same consensus position in the alignment is called a node. The states are interconnected by arrows as shown in Figure 1, representing state

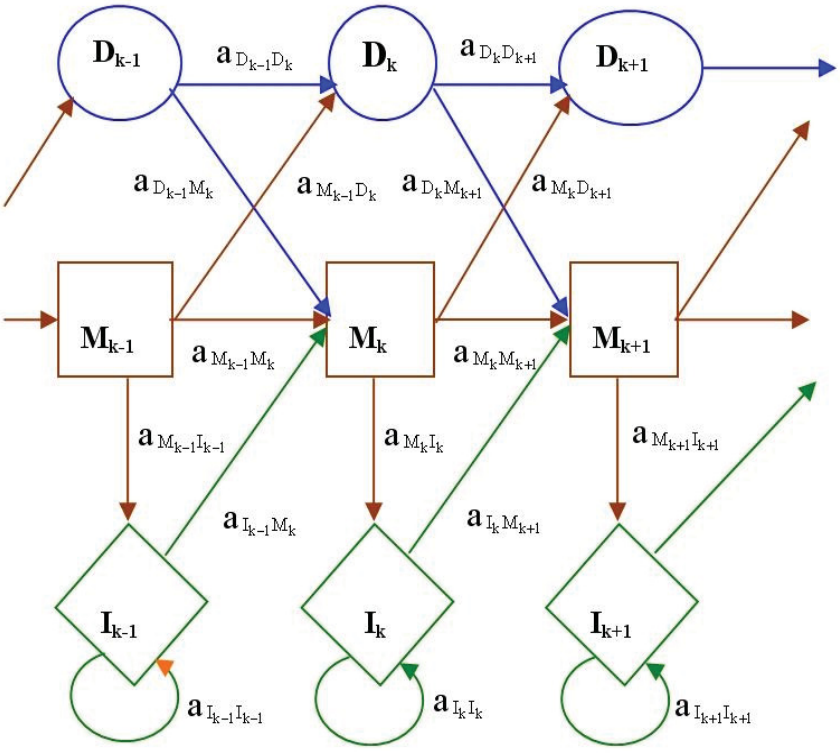


**Fig. 1** Profile HMM architecture based on Plan 7 (HMMER 2).

transition probabilities. The transitions are arranged so that at each node, either an M state is triggered (a residue is aligned and scored) or a D state is triggered (no residue is aligned, resulting in a deletion-gap character "–"). Insertions occur between nodes, and an I state can have a self-transition, allowing one or more inserted residues to occur between consensus columns. The transition to an I state for the first inserted residue, followed by zero or more I→I self-transitions for each subsequent inserted residue, is the probabilistic equivalent of the familiar gap-open and gap-extend affine gap penalty system. Like all HMMs, profile HMMs have emission and transition probabilities with probability distribution over the whole space of sequences, which is parameterized using Baum-Welch re-estimation formulas to peak the distribution around the members of the family (*5*). Forward algorithm, backward algorithm, re-estimation algorithm, and Viterbi algorithm are the four main components of profile HMM.

***Forward algorithm***

Forward algorithm is used to calculate the log-odd scores of a protein sequence. The forward variables for classical profile HMM, namely, $f_{M_k}(i)$, $f_{I_k}(i)$, and $f_{D_k}(i)$ for the $k^{\text{th}}$ match, insert, and delete state are estimated using Equations 1–3, respectively (*5*):

$$\begin{aligned} f_{M_k}(i) &= e_{M_k}(i)[f_{M_{k-1}}(i-1)a_{M_{k-1}M_k} \\ &\quad + f_{I_{k-1}}(i-1)a_{I_{k-1}M_k} \\ &\quad + f_{D_{k-1}}(i-1)a_{D_{k-1}M_k}] \qquad (1) \\ f_{I_k}(i) &= e_{I_k}(i)[f_{M_k}(i-1)a_{M_kI_k} + f_{I_k}(i-1)a_{I_kI_k} \\ &\quad + f_{D_k}(i-1)a_{D_kI_k}] \qquad (2) \\ f_{D_k}(i) &= f_{M_{k-1}}(i)a_{M_{k-1}D_k} + f_{I_{k-1}}(i)a_{I_{k-1}D_k} \\ &\quad + f_{D_{k-1}}(i)a_{D_{k-1}D_k}] \qquad (3) \end{aligned}$$

***Backward algorithm***

Backward algorithm is used for parameter estimation. In classical profile HMM, the backward variables $b_{M_k}(i)$, $b_{I_k}(i)$, and $b_{D_k}(i)$ for the $k^{\text{th}}$ match, insert, and delete state, respectively, are calculated as shown in Equations 4–6 (*5*):

$$\begin{aligned} b_{M_k}(i) &= [b_{M_{k+1}}(i+1)\, a_{M_kM_{k+1}}\, e_{M_{k+1}}(x_{i+1}) \\ &\quad + b_{I_k}(i+1)\, a_{M_kI_k}\, e_{I_{k+1}}(x_{i+1}) \\ &\quad + b_{D_{k+1}}(i)\, a_{M_kD_{k+1}}] \qquad (4) \\ b_{I_k}(i) &= [b_{M_{k+1}}(i+1)\, a_{I_kM_{k+1}}\, e_{M_{k+1}}(x_{i+1}) \\ &\quad + b_{I_k}(i+1)\, a_{I_kI_k}\, e_{I_{k+1}}(x_{i+1}) \\ &\quad + b_{D_{k+1}}(i)\, a_{I_kD_{k+1}}] \qquad (5) \\ b_{D_k}(i) &= [b_{M_{k+1}}(i+1)\, a_{D_kM_{k+1}}\, e_{M_{k+1}}(x_{i+1}) \\ &\quad + b_{I_k}(i+1)\, a_{D_kI_k}\, e_{I_k}(x_{i+1}) \\ &\quad + b_{D_{k+1}}(i)\, a_{D_kD_{k+1}}] \qquad (6) \end{aligned}$$

***Re-estimation algorithm***

The emission and transition matrices for classical profile HMM are re-estimated by computing all the elements of the emission and transition matrices as shown in Equations 7–11:

$$E_{M_k}(a) = \frac{1}{P(O)} \sum_{i|O_i=a} f_{M_k}(i) b_{M_k}(i) \qquad (7)$$

$$E_{I_k}(a) = \frac{1}{P(O)} \sum_{i|O_i=a} f_{I_k}(i) b_{I_k}(i) \qquad (8)$$

$$\begin{aligned} A_{M_kM_{k+1}} &= \frac{1}{P(O)} \sum_i f_{M_k}(i) a_{M_kM_{k+1}} e_{M_{k+1}}(x_{i+1}) \\ &\quad b_{M_{k+1}}(i+1) \qquad (9) \\ A_{M_kI_k} &= \frac{1}{P(O)} \sum_i f_{M_k}(i) a_{M_kI_k} e_{I_k}(x_{i+1}) \\ &\quad b_{I_k}(i+1) \qquad (10) \\ A_{M_kD_{k+1}} &= \frac{1}{P(O)} \sum_i f_{M_k}(i) a_{M_kD_{k+1}} b_{D_{k+1}}(i) \qquad (11) \end{aligned}$$

$P(O)$ represents the probability of the sequence.

***Viterbi algorithm***

Classical Viterbi algorithm is used to compute the negative logarithm of the probability of the single most likely path, $\widehat{\delta}$, for a given sequence $O$. It can be formulated as:

$$\widehat{\delta} = -\log \max_{\widehat{\pi}} P(O|\widehat{\pi}, \widehat{\lambda}) \qquad (12)$$

where $\widehat{\pi}$ represents the path containing the sequence of states (M, I, and D) that emitted the amino acid residues in sequence $O$ for the given model $\widehat{\lambda}$. In classical profile HMM, the Viterbi variables $\widehat{\delta}_{M_k}(i)$, $\widehat{\delta}_{I_k}(i)$, and $\widehat{\delta}_{D_k}(i)$ for the $k^{\text{th}}$ match, insert, and delete state, respectively, are calculated as shown in Equations 13–15 (*5*):

$$\widehat{\delta}_{M_k}(i) = -\log(\max\{e_{M_k}(i)\,\widehat{\delta}_{M_{k-1}}(i-1)\,a_{M_{k-1}M_k}, e_{M_k}(i)\,\widehat{\delta}_{I_{k-1}}(i-1)\,a_{I_{k-1}M_k}, e_{M_k}(i)\,\widehat{\delta}_{D_{k-1}}(i-1)\,a_{D_{k-1}M_k}\}) \quad (13)$$

$$\widehat{\delta}_{I_k}(i) = -\log(\max\{e_{I_k}(i)\,\widehat{\delta}_{M_k}(i-1)\,a_{M_kI_k}, \widehat{\delta}_{I_k}(i-1)\,a_{I_kI_k}, \widehat{\delta}_{D_k}(i-1)\,a_{D_kI_k}\}) \quad (14)$$

$$\widehat{\delta}_{D_k}(i) = -\log(\max\{\widehat{\delta}_{M_{k-1}}(i)\,a_{M_{k-1}D_k}, \widehat{\delta}_{I_{k-1}}(i)\,a_{I_{k-1}D_k}, \widehat{\delta}_{D_{k-1}}(i)\,a_{D_{k-1}D_k}, D_k)\}) \quad (15)$$

The equations formulated above are based on the Plan 9 architecture (*15*). They can be easily extended to Plan 7 architecture by setting the transition parameters $a_{I_{k-1}D_k}$ and $a_{D_kI_k}$ to zero. Since protein sequences have high degrees of interdependencies, the additive hypothesis of probability measure is not well suited. The classical model (HMMER) based on probability theory assigns the same level of importance to the source, that is, the states in profile HMM. A more flexible way to overcome this limitation is provided by fuzzy measures and integrals (*16*). They take into account the relative importance of the source along with the information (*8*).

## Fuzzy measures and integrals

Probability measure theory obeys the additivity of classical theory by assigning one to the universal set. Fuzzy measures are an extension of the classical additive theory. They are obtained by replacing the additive requirement of classical measures with weaker properties of monotonicity, continuity, and semi-continuity (*8*). The aggregation of fuzzy measures is done using Choquet or Sugeno integrals.

### *Fuzzy measure*

Let $\Omega$ be the power set of a universal set $X$. A set function $g : \Omega \rightarrow [0,1]$ defined on $\Omega$, which satisfies the conditions of boundary, monotonicity, and continuity shown in Equations 16–18, is called a fuzzy measure. It represents the mapping of a crisp power set of a universal set to a unit interval representing evidence.

$$\text{Boundary: } g(\phi) = 0,\ g(X) = 1 \quad (16)$$

$$\text{Monotonicity: If A, B} \subseteq \Omega \text{ and A} \subseteq \text{B, then g(A)} \leq \text{g(B)} \quad (17)$$

Continuity: For any increasing sequence $A_1 \subseteq A_2 \subseteq \cdots \subseteq A_i \cdots$ of sets in $\Omega$, if $\bigcup_{i=1}^{\infty} A_i \in \Omega$, then

$$\lim_{i\rightarrow\infty} \text{g(A}_i\text{)} = \text{g}(\bigcup_{i=1}^{\infty} A_i) \quad (18)$$

From the definition of a fuzzy measure $g$, the union of two disjoint subsets cannot be directly computed from the component measures. Possibility measure based on t-conorm S is one of the most widely used fuzzy measures. The t-conorm S is an operation on the unit interval [0,1] satisfying the following conditions on elements a, b, and c (*17*).

$$\text{Neutrum element : S(a, 0)} = \text{a} \quad (19)$$

$$\text{Monotonicity : b} \leq \text{c} \rightarrow \text{S(a, b)} \leq \text{S(a, c)} \quad (20)$$

$$\text{Commutativity : S(a, b)} = \text{S(b, a)} \quad (21)$$

$$\text{Associativity : S}\big(\text{a, S(b, c)}\big) = \text{S}\big(\text{S(a, b), c}\big) \quad (22)$$

Two types of S operations, maximum and drastic t-conorm operators, are shown in Equations 23 and 24, respectively:

$$\text{S(a, b)} = max(\text{a, b}) \text{ if a and b} \neq 0 \quad (23)$$

$$\text{S(a, b)} = \begin{cases} \text{a} & \text{if b} = 0, \\ \text{b} & \text{if a} = 0, \\ 1 & \text{if a and b} \neq 0 \end{cases} \quad (24)$$

Possibility measure is based on the above defintions of max t-conorm operation. If $X$ is a universal set with $\Omega$ consisting of all the subsets of $X$, then the possibility measure $g_P$ is:

$$g_P : \Omega \rightarrow [0,1] \quad (25)$$

where $g_P(\phi)= 0$ and $g_P(X) = 1$. It satisfies the constraints shown in Equation 26 along with the ones defined above for the fuzzy measures.

$$g_P(\bigcup_{A_i \in \Omega} A_i) = max\big(g_P(A_i)\big) \quad (26)$$

The possibility measures on each element of the set $X$ denoted by $g_P(x)$ are called the *possibility-density* measures. Using these density measures, we can calculate the possibility measures for all the sets in $\Omega$:

$$g_P(A_i) = max\big(g_P(x)\big),\ \forall\ x \in A_i \quad (27)$$

***Fuzzy integrals***

Fuzzy integrals, defined with respect to fuzzy measures, are nonlinear functions combining multiple sources of uncertain information (*7*). They use information concerning the importance of individual source as well as source subsets to derive a reasonable numerical confidence value for the particular hypothesis decision under consideration. Here, we give a brief description of the Choquet integral, which is one of the most commonly used fuzzy integrals.

Let $(X, \Omega)$ be a measurable space and let $h : X \rightarrow [0, 1]$ be an $\Omega$ measurable function. The Choquet integral over $A \subseteq X$ of the function $h$ with respect to a fuzzy measure $g$ is defined by:

$$\int_X h(x) \circ g(\cdot) = \int_0^1 g(A_\alpha) d\,\alpha \tag{28}$$

where $A_\alpha = \{x | h(x) > \alpha\}$. For a discrete set $X$ with $N$ elements, the Choquet integral $(e_c)$ can be computed as follows:

$$e_c = \sum_{i=1}^{N} h(x_i)[g(A_i) - g(A_{i+1})] \tag{29}$$

where $h(x_1) \leq h(x_2) \leq \cdots \leq h(x_N)$ and $A_i = \{x_i, x_{i+1}, \cdots, x_N\}$.

## Fuzzy profile HMM

In classical profile HMM, the joint probability measure $P(O_1, O_2, \cdots, O_t, O_{t+1}, q_{t+1} = Z_j)$ is written as the product $P(O_1, O_2, \cdots, O_t, O_{t+1} | q_{t+1} = Z_j) \cdot P(q_{t+1} = Z_j)$, thus making the following two assumptions of the statistical independence.

- The amino acid $O_{t+1}$ emitted by the HMM at position $t + 1$ at $Z_j$ state is independent of the previously emitted amino acid sequences $(O_1, O_2, \cdots, O_t)$ (*6*).
- The active state at position $t + 1$, $Z_j$, is independent of the previous subsequence of amino acids $(O_1, O_2, \cdots, O_t)$ observed.

These assumptions are not realistic for the homologous sequences of a family since they have a high correlation among neighboring residues (*3*). Improved results for building profiles can be expected through the relaxation permitted by fuzzy measures leading to the fuzzy profile HMM. As mentioned above, in the fuzzy profile HMM, the additive property of probability measures is replaced with the weaker condition of monotonicity by using fuzzy measures and integrals (*8*). The Choquet integral, used to aggregate the fuzzy measures, takes into account the importance of the individual and subsets of source (states and subset of states). The fuzzy forward and backward variables form the basis of the fuzzy Viterbi algorithm used for alignments and the fuzzy Baum-Welch algorithm for parameter estimation (*12*). The fuzzy profile HMM, $\overline{\theta} = (\widehat{A}, \widehat{B}, \widehat{\pi})$, is characterized by the following parameters (*12*).

| | |
|---|---|
| O | protein sequence |
| T | length of the protein sequence |
| N | profile model length |
| $\Omega$ | set of protein sequences of the family |
| X | finite set of states at position $t$ |
| Y | finite set of states at position $t + 1$ |
| Z | states $\{Z_1, Z_2, \cdots, Z_N\}$ |
| $\widehat{\pi}_Z(\cdot)$ | initial state fuzzy measure |
| $\widehat{\pi}_Z(\{Z_i\})$ | initial state fuzzy density |
| $\widehat{b}_j(O_t)$ | symbol fuzzy density |
| $\widehat{a}_y(\cdot \mid X)$ | transition fuzzy measure |
| $\widehat{a}_y(y_j \mid x_i)$ | transition fuzzy density |
| $q_t$ | state visited at position $t$ |

where $\widehat{A} = [\widehat{a}_y(y_j|x_i) = \widehat{a}_{ij}]$, $\widehat{B} = [\widehat{b}_j(O_t)]$, and $\widehat{\pi} = [\widehat{\pi}_Z(\{Z_i\})]$.

***Fuzzy forward algorithm***

We formulate the fuzzy forward variable, $\widehat{f}_{\Omega_y}(\{O_1, O_2, \cdots, O_t\} \times \{y_j\})$ for the fuzzy profile HMM, which can be reduced to the combination of two measures defined on $\{O_1, O_2, \cdots, O_t\}$ and on the states $y_j = (M_{t+1}, I_t, D_{t+1})$. This avoids the assumption of decomposition of measures as done in classical HMMs. At any time, the fuzzy measure $\widehat{f}_{\Omega_y}$ on $\Omega_{1,t+1} \times Y$ can be constructed from its constituent forward variables through recursion, after integrating with the Choquet integral and with multiplication as an intersection operator. This is shown by the following equations:

$$f_{y_j}(t+1) = \widehat{f}_{\Omega_y}(\{O_1, O_2, \cdots, O_t\} \times \{y_j\}) \tag{30}$$

$$= \int_X \widehat{a}_y(\{y_j\}|x) \circ \widehat{f}_{\Omega_X}(\{O_1, O_2, \cdots, O_t\}) \wedge \widehat{b}_j(O_{t+1}) \tag{31}$$

where $\wedge$ is the fuzzy intersection operator and $\circ$ is the multiplication operator. The elements of matrix $\widehat{A} = [\widehat{a}_y(y_j|x_i)]$, containing the probability values for

transition to state $y_j$ from state $x_i$, are assigned accordingly to function $h$ as shown below:

$$h(x_i, y_j) = \widehat{a}_y(y_j|x_i) \tag{32}$$

All the values $h(x_i, y_j)$ representing the transition probabilities to state $y_j$ are sorted in Equation 33:

$$h(x_1, y_j) \leq h(x_2, y_j) \leq \ldots \leq h(x_i, y_j) \leq \ldots \leq h(x_N, y_j) \tag{33}$$

Based on the above sorting, a set $k_i(y_j)$ is obtained as:

$$k_i(y_j) = \{x_i, x_{i+1}, \ldots, x_N\} \tag{34}$$

where $x_i$ is the state number at the $i^{\text{th}}$ position according to constraints in Equation 33 based on transition to the ${y_j}^{\text{th}}$ state from all other states. According to the definition of fuzzy measures and fuzzy integrals, $f_{y_j}(t+1)$ is given by Equation 35, which satisfies the constraints in Equations 33 and 34.

$$\begin{aligned} f_{y_j}(t+1) &= \sum_{i=1}^{N} h(x_i, y_j)[g(k_i(y_j)) - g(k_{i+1}(y_j))] \\ &= \sum_{i=1}^{N} h(x_i, y_j) d_i(y_j) \end{aligned} \tag{35}$$

where $d_i(y_j)$ represents the difference between successive fuzzy measures and $g(k_i(y_j))$ represents the fuzzy measure. After normalizing the difference between successive fuzzy measures with respect to fuzzy density $f_{x_i}(t)$, we obtain:

$$\rho_t(x_i, y_j) = d_i(y_j)/f_{x_i}(t) \tag{36}$$

Based on Equations 35 and 36, the forward variable for the ${y_j}^{\text{th}}$ state at position $t+1$ can be reformulated as:

$$f_{y_j}(t+1) = \sum_{i=1}^{N} h(x_i, y_j)\rho_t(x_i, y_j) f_{x_i}(t) \circ \widehat{b}_j(O_{t+1}) \tag{37}$$

Accordingly, we reformulate the forward variables $f_{M_k}(i), f_{I_k}(i)$, and $f_{D_k}(i)$ for the $k^{\text{th}}$ match, insert, and delete state, respectively, using the possibility measure as shown below:

$$\begin{aligned} f_{M_k}(i) &= e_{M_k}(i)[f_{M_{k-1}}(i-1)a_{M_{k-1}M_k}\rho_{i-1}(M_{k-1}, M_k) \\ &\quad + f_{I_{k-1}}(i-1)a_{I_{k-1}M_k}\rho_{i-1}(I_{k-1}, M_k) \\ &\quad + f_{D_{k-1}}(i-1)a_{D_{k-1}M_k}\rho_{i-1}(D_{k-1}, M_k)] \end{aligned} \tag{38}$$

$$\begin{aligned} f_{I_k}(i) &= e_{I_k}(i)[f_{M_k}(i-1)a_{M_k I_k}\rho_{i-1}(M_k, I_k) \\ &\quad + f_{I_k}(i-1)a_{I_k I_k}\rho_{i-1}(I_k, I_k) \\ &\quad + f_{D_k}(i-1)a_{D_k I_k}\rho_{i-1}(D_k, I_k)] \end{aligned} \tag{39}$$

$$\begin{aligned} f_{D_k}(i) &= f_{M_{k-1}}(i)a_{M_{k-1}D_k}\rho_{i-1}(M_{k-1}, D_k) \\ &\quad + f_{I_{k-1}}(i)a_{I_{k-1}D_k}\rho_{i-1}(I_{k-1}, D_k) \\ &\quad + f_{D_{k-1}}(i)a_{D_{k-1}D_k}\rho_{i-1}(D_{k-1}, D_k)] \end{aligned} \tag{40}$$

The term $\rho$ in the above equations represents the fuzzy measure difference, which is calculated using the Choquet integral as shown in Algorithm 1.

**Algorithm 1**: Calculation of fuzzy measure difference density-$\rho$

**input** : $\widehat{\chi} = (\widehat{A}, \widehat{B}, \widehat{\pi})$
**output** : $\rho_t(x_i, y_j)\ \forall\ y_j \in Y$ and $x_i \in X$

*Repeat for the length = T of the sequence*;
**for** $t$ = 1 ***to*** $T$ **do**
  *Evaluate fuzzy densities for each state x at position t which is a measure of importance of each state*;
  **for** $i$ = 1 ***to*** $N$ *and* $x_i \in X$ **do**
    $g(\{x_i\}) = f_{x_i}(t)$;
  **end**
  *Define fuzzy measures sets corresponding to transition to state* $y_j$ *at position t+1*;
  **for** $j$ = 1 ***to*** $N$ *and* $y_j \in Y$ **do**
    *Set containing the indices of the states*;
    $g(k_i(y_j)) = g(\{x_i, x_{i+1}, \ldots, x_N\})$;
    *Calculate the possibility fuzzy measures based on the sets*;
    $g(k_i(y_j)) = \wedge(g(\{x_i\}), g(k_{i+1}(y_j)))$;
    **for** $i$ = 1 ***to*** $N$ **do**
      *Calculate* $\rho_t(i, y_j)$ ;
      $d_i(y_j) = g(k_i(y_j)) - g(k_{i+1}(y_j))$, $\rho_t(x_i, y_j) = d_i(y_j)/g(\{x_i(y_j)\})$;
    **end**
  **end**
**end**

***Fuzzy backward algorithm***

The fuzzy backward variable $b_{s_i}(t+1)$ is a conditional fuzzy measure, measuring the fuzziness of the observation $O_1, O_2, \cdots, O_t$ because of visiting state $Z_i$.

$$b_{Z_i}(t) = \widehat{b}_{\Omega_{t-1,T}}(\{O_{t+1}, \cdots, O_T\} \times \{x_i\}) \quad (41)$$

Equation 42 shows the fuzzy backward variable when integrated using the Choquet integral with respect to any fuzzy measure and multiplication as the intersection operator.

$$b_{s_i}(t) = \int_Y [\widehat{b}_{\Omega_{t-2,T}}(\{O_{t+1}, \cdots, O_T\}|y) \wedge \widehat{b}_j(O_{t+1})] \circ \widehat{a}_Y(\cdot|x) \quad (42)$$

We reformulate the backward variables $b_{M_k}(i)$, $b_{I_k}(i)$, and $b_{D_k}(i)$ for the $k^{\text{th}}$ match, insert, and delete state, respectively, using the possibility measure as shown below:

$$\begin{aligned} b_{M_k}(i) = {} & b_{M_{k+1}}(i+1) a_{M_k M_{k+1}} \rho_{i+1}(M_k, M_{k+1}) \\ & e_{M_{k+1}}(i+1) + b_{I_k}(i+1) a_{M_k I_k} \rho_{i+1}(M_k, I_k) \\ & e_{I_{k+1}}(i+1) + b_{D_{k+1}}(i) a_{M_k D_{k+1}} \rho_i(M_k, D_{k+1}) \end{aligned} \quad (43)$$

$$\begin{aligned} b_{I_k}(i) = {} & b_{M_{k+1}}(i+1) a_{I_k M_{k+1}} \rho_{i+1}(I_k, M_{k+1}) \\ & e_{M_{k+1}}(i+1) + b_{I_k}(i+1) a_{I_k I_k} \rho_{i+1}(I_k, I_k) \\ & e_{I_{k+1}}(i+1) + b_{D_{k+1}}(i) a_{I_k D_{k+1}} \rho_i(I_k, D_{k+1}) \end{aligned} \quad (44)$$

$$\begin{aligned} b_{D_k}(i) = {} & b_{M_{k+1}}(i+1) a_{D_k M_{k+1}} \rho_{i+1}(D_k, M_{k+1}) \\ & e_{M_{k+1}}(i+1) + b_{I_k}(i+1) a_{D_k I_k} \rho_{i+1}(D_k, I_k) \\ & e_{I_k}(i+1) + b_{D_{k+1}}(i) a_{D_k D_{k+1}} \rho_i(D_k, D_{k+1}) \end{aligned} \quad (45)$$

***Fuzzy Baum-Welch re-estimation algorithm***

After formulating the forward and backward variables, we extend fuzzification to parameter estimation methods for the profile HMM. The emission and transition matrices for the fuzzy profile HMM are re-estimated by computing all the elements of the emission and transition matrices as given by Equations 46–50:

$$E_{M_k}(a) = \frac{1}{P(O)} \sum_{i|O_i=a} f_{M_k}(i) \rho_i(M_k, \cdot) b_{M_k}(i) \quad (46)$$

$$E_{I_k}(a) = \frac{1}{P(O)} \sum_{i|O_i=a} f_{I_k}(i) \rho_i(I_k, \cdot) b_{I_k}(i) \quad (47)$$

$$\begin{aligned} A_{M_k M_{k+1}} = {} & \frac{1}{P(O)} \sum_i f_{M_k}(i) a_{M_k M_{k+1}} \rho_i(M_k, M_{k+1}) \\ & e_{M_{k+1}}(x_{i+1}) b_{M_{k+1}}(i+1) \end{aligned} \quad (48)$$

$$\begin{aligned} A_{M_k I_k} = {} & \frac{1}{P(O)} \sum_i f_{M_k}(i) a_{M_k I_k} \rho_i(M_k, I_k) \\ & e_{I_k}(x_{i+1}) b_{I_k}(i+1) \end{aligned} \quad (49)$$

$$\begin{aligned} A_{M_k D_{k+1}} = {} & \frac{1}{P(O)} \sum_i f_{M_k}(i) a_{M_k D_{k+1}} \rho_i(M_k, D_{k+1}) \\ & b_{D_{k+1}}(i) \end{aligned} \quad (50)$$

The transition parameters for insert and delete states can be calculated similarly.

***Fuzzy Viterbi algorithm***

The classical Viterbi algorithm can be modified using fuzzy measures to compute $\widehat{\delta}_{Z_i}(t)$ at position $t$ for state $Z_i$ as shown below (7):

$$\widehat{\delta}_{Z_i}(t) = \max_{q_1, q_2, \cdots, q_{t-1}} P(q_1, q_2, \cdots, q_t = Z_i, O_1, O_2, \cdots, O_t|\widehat{\lambda}) \quad (51)$$

where $q_t$ represents the state visited at position $t$, emitting amino acid residue $O_t$, and can be either match, insert, or delete state represented by $Z_i$. The maximization is modified using fuzzy measure difference density $\rho$ to obtain fuzzy Viterbi algorithm:

$$\begin{aligned} \widehat{\delta}_{Z_i}(t) = {} & \max_{q_1, q_2, \cdots, q_{t-1}} \{\widehat{\pi}_{q_1} \widehat{b}_{q_1}(O_1) \prod_{\tau=2}^{t} [\widehat{a}_{q_{\tau-1} q_\tau} \rho_\tau(q_{\tau-1}, \\ & q_\tau)] \widehat{b}_{q_\tau}(O_\tau)\} \end{aligned} \quad (52)$$

where $\widehat{\pi}_{q_1}$ represents the initial state fuzzy density, and $\rho_\tau(q_{\tau-1}, q_\tau)$ represents the fuzzy measure difference density. $\widehat{\delta}_{Z_i}(t+1)$ is computed recursively for the entire length of the sequence as shown in Algorithm 2.

Based on Equation 1 and Algorithm 1, $\widehat{\delta}_{M_k}(i)$, $\widehat{\delta}_{I_k}(i)$, and $\widehat{\delta}_{D_k}(i)$ for the $k^{\text{th}}$ match, insert, and delete state, respectively, can be formulated as shown below:

$$\begin{aligned} \widehat{\delta}_{M_k}(i) = {} & -\log(\max\{e_{M_k}(i) \widehat{\delta}_{M_{k-1}}(i-1) a_{M_{k-1} M_k} \\ & \rho_{i-1}(M_{k-1}, M_k), e_{M_k}(i) \widehat{\delta}_{I_{k-1}}(i-1) \\ & a_{I_{k-1} M_k} \rho_{i-1}(I_{k-1}, M_k), e_{M_k}(i) \widehat{\delta}_{D_{k-1}}(i-1) \\ & a_{D_{k-1} M_k} \rho_{i-1}(D_{k-1}, M_k)\}) \end{aligned} \quad (53)$$

$$\widehat{\delta}_{I_k}(i) = -\log(\max\{e_{I_k}(i)\widehat{\delta}_{M_k}(i-1)a_{M_kI_k} \rho_{i-1}(M_k,I_k), \widehat{\delta}_{I_k}(i-1)a_{I_kI_k}\rho_{i-1}(I_k,I_k), \widehat{\delta}_{D_k}(i-1)a_{D_kI_k}\rho_{i-1}(D_k,I_k)\}) \quad (54)$$

$$\widehat{\delta}_{D_k}(i) = -\log(\max\{\widehat{\delta}_{M_{k-1}}(i)a_{M_{k-1}D_k} \rho_{i-1}(M_{k-1},D_k), \widehat{\delta}_{I_{k-1}}(i)a_{I_{k-1}D_k} \rho_{i-1}(I_{k-1},D_k), \widehat{\delta}_{D_{k-1}}(i)a_{D_{k-1}D_k} \rho_{i-1}(D_{k-1},D_k)\}) \quad (55)$$

**Algorithm 2**: Fuzzy Viterbi algorithm

*Initialization for N match, insert and delete states*;

**for** $1 \le i \le N$ **do**

$\widehat{\delta}_{Z_i}(1) = \widehat{\pi}_{Z_i}\widehat{b}_{Z_i}(O_1)$

$\widehat{\psi}_{Z_i}(1) = 0$

**end**

*Recursion through the length of the sequence T for N states*;

**for** $2 \le \tau \le T$ **do**

**for** $1 \le j \le N$ **do**

$\widehat{\delta}_{Z_j}(\tau) = \max_{1\le i\le N}[\widehat{\delta}_{Z_i}(\tau-1)\widehat{a}_{z_iz_j}\rho_\tau(z_i,z_j)]\widehat{b}_{Z_j}(O_\tau)$

$\widehat{\psi}_{Z_j}(\tau) = \arg\max_{1\le i\le N}[\widehat{\delta}_{Z_i}(\tau-1)\widehat{a}_{z_iz_j}\rho_\tau(z_i,z_j)]$

**end**

**end**

*Termination conditions*;

$\widehat{P}^* = \max_{1\le i\le N}[\widehat{\delta}_{Z_i}(T)] \quad \widehat{q}^* = \arg\max_{1\le i\le N}[\widehat{\delta}_{Z_i}(T)]$

*Backtracking for optimal paths*;

**for** $1 \le \tau \le T\text{-}1$ **do**

$\widehat{q}^*_\tau = \widehat{\psi}_{\widehat{q}^*_{\tau+1}}(\tau+1)$

**end**

### *Computational complexity analysis*

In a classical (HMMER) profile model with $N$ states, the forward variables and the Viterbi algorithm have a computational complexity of the order of $O(N^2T)$ in time (*15*) for a protein sequence of length $T$. At any instant of time, transitions occur to the $k^{\text{th}}$ M state only from the $(k-1)^{\text{th}}$ M, I, and D states. Similarly, D and I states also have only three incoming transitions that reduce the computational complexity to $O(3^NT)$. In the fuzzy profile model, the computational complexity is of the order of $O\big((2^{N-1})^NT\big)$ since $2^{N-1}$ subsets are computed at each state during the forward variable calculation. The computational complexity for the fuzzy profile model can be reduced to $O(7^NT)$. The fuzzy model is computationally expensive compared with the classical model, but the trade-off is provided by improved accuracy of family identification and biologically significant alignments. As the primary goal is to improve the accuracy, the issue of computational complexity becomes secondary, since these computations are carried offline.

# Evaluation

We evaluated the performance of the fuzzy profile HMM using sequences of widely studied globin and kinase families, and compared the results with those of the HMMER profile model.

## Evaluation on globins

Globins are part of a large family of heme-containing proteins involved in the storage and transport of oxygen that have different oligomeric states and overall architecture (*18*, *19*). They are responsible for binding and/or transporting oxygen. The major groups of globins are hemoglobins and myoglobins from vertebrates and invertebrates, leghemoglobins from plants, and flavohemoglobins from bacteria. Hemoglobin is a protein responsible for transporting oxygen from the lungs to other tissues, and is a tetramer of two $\alpha$ and $\beta$ chains each. We extacted the globin sequences from the SWISS-PROT database (*20*) by searching the keyword "globin". The globin dataset sample used in the evaluation consists of 625 different globin sequences. These sequences also belong to the Pfam (*21*, *22*) domains with accession numbers PF00042, PF0152, PF01099, and PF06438. The sequences vary in length from 109 to 428 amino acids.

The globin dataset was divided into 12 random folds. The model parameters were trained and optimized using one of the folds and the remaining folds were used as test dataset. To incorporate noise into the model, 1,953 non-globin sequences were added to the test dataset. The non-globin sequences varied in length from 25 to 350 amino acids and were obtained from SWISS-PROT database. The match of sequences to classical and fuzzy profile HMMs was scored using log-odd scores (defined later). The alignments for globin sequences were obtained through fuzzy and classical Viterbi algorithms. The classical Viterbi algorithm was used to align the sequences to the globin profile model based on HMMER. The estimation of both fuzzy and classical model parameters was done 12 times and the models with the highest overall log-likelihood scores were selected. Similar

cross-fold validations have been carried out in earlier studies (*5*).

The performance of the fuzzy model was evaluated and compared with the HMMER profile model based on estimation comparison and Z-score plots.

### *Estimation comparison*

The transition parameters were obtained using the classical and fuzzy Baulm-Welch re-estimation methods on the 50 globin sequences used for training. Figures 2 and 3 graphically depict the converged transition probabilities for all match states of classical and fuzzy profile HMMs, respectively. The transition probability from the $k^{\text{th}}$ match state to the $(k+1)^{\text{th}}$ match, the $k^{\text{th}}$ insert, and the $(k+1)^{\text{th}}$ delete state are represented by the top, middle, and bottom subgraphs, respectively. It is observed from the transition diagrams that the classical and fuzzy HMMs learn different values at some specific points, indicating the difference in their observed behavior. We further observed the following behaviors for transition matrices in classical and fuzzy profile HMMs.

1. The classical profile HMM has 5 transition probability values from the $k^{\text{th}}$ match state to the $k^{\text{th}}$ insert state with value greater than 0.05 before the state number 60 is reached (Figure 2). This indicates that the classical HMM has more insertions compared with the fuzzy profile model, which has only 3 transition probability values greater than 0.05 in the same region (Figure 3).

2. The more insert transitions observed in the classical HMM compared with the fuzzy HMM indicates that there are more transition probabilities from the $k^{\text{th}}$ insert state to the $(k+1)^{\text{th}}$ match state.

3. Both classical and fuzzy profile models have the same nature for the transitions from delete states.

The emission parameters were also obtained using the classical and fuzzy Baum-Welch re-estimation methods on the 50 globin sequences used for training. Figures 4 and 5 show the emission probability distribution of 20 amino acids at different match states (25, 50, 100, and 125) obtained by classical and fuzzy models, respectively. When match state M = 125, the residue histidine has the highest emission probability according to the classical model, while serine is the highest in the fuzzy profile model. This difference in emission distributions contributes to a different consensus alignment.

### *Z-score plot*

Log-odd score is given by the ratio of most probable alignment $\Pi^*$ of the sequence $R$ for a given model $\lambda$ with respect to the probability of $R$ through a null random model ($\Upsilon$):

$$\text{Log-odd score} = \frac{P(R\ \Pi^*, |\lambda)}{P(R|\Upsilon)} \tag{56}$$

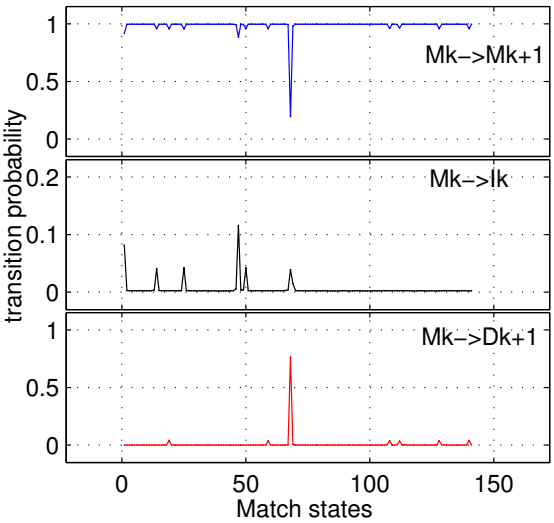


**Fig. 2** Distribution of match (M) state transitions for the classical profile HMM trained by the classical Baum-Welch estimation algorithm.

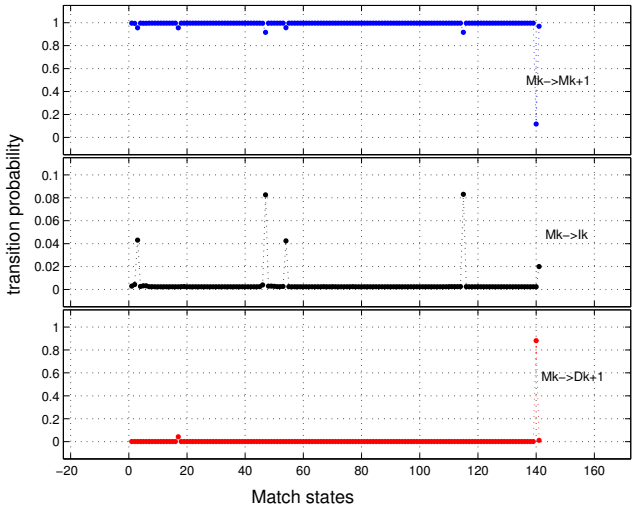


**Fig. 3** Distribution of match (M) state transitions for the fuzzy profile HMM trained by the fuzzy Baum-Welch estimation algorithm.

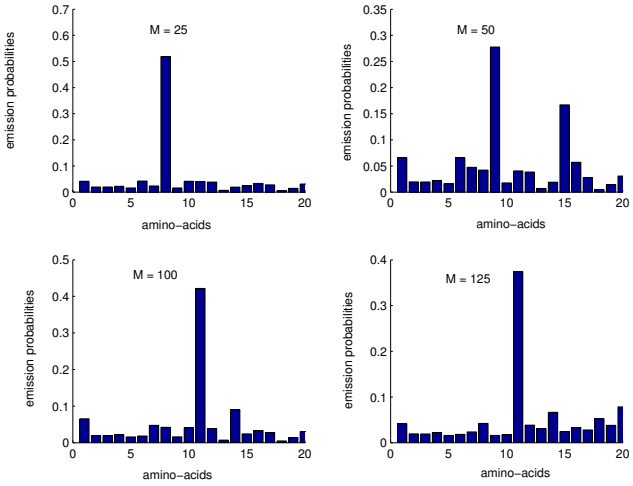


**Fig. 4** Emission distribution of 20 amino acids for match states (25, 50, 100, 125) of the classical profile HMM trained by the classical Baum-Welch estimation algorithm.

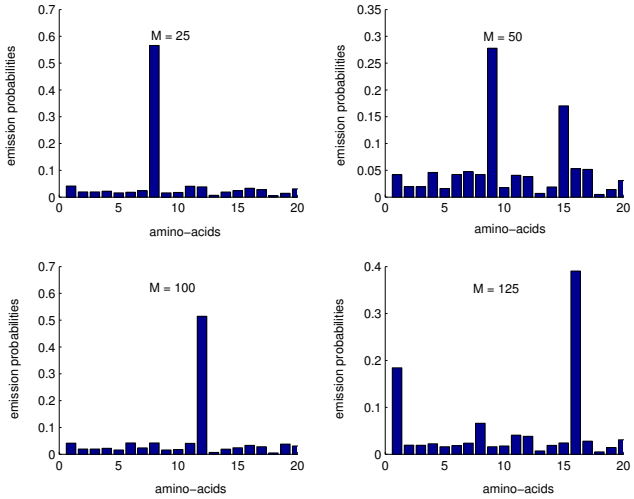


**Fig. 5** Emission distribution of 20 amino acids for match states (25, 50, 100, 125) of the fuzzy profile HMM trained by the fuzzy Baum-Welch estimation algorithm.

The probability of $R$ through $RM$, which assumes that the underlying sequences are unrelated, is given by the simple product of frequencies of residues as shown below:

$$P(R|\Upsilon) = \prod_{i=1}^{L} \left(e_{\Upsilon}(R_i)\right) \tag{57}$$

This ratio provides a significance assessment of log-odd scores (*5*). The amino acid frequencies of the sequences in training dataset are used for $\Upsilon$. To calculate Z-score, a smooth curve is fitted for the log-odd score plot using the local window technique (*5*). A standard deviation is estimated for each length and Z-score is calculated for each score by estimating its distance from the curve in terms of standard deviation. The normalized Z-score plots for the classical (HMMER) and fuzzy profile models are shown in Figures 6 and 7, respectively. For the fuzzy model, it is observed that all the globin member sequences (both from training and test datasets) are clustered with the normalized Z-score between 2.0 to 8.0. This is mainly because of the *max* operation performed by the possibility measure. The member sequences are scattered sparsely with the normalized Z-score varying from 1.0 to 8.0 for the HMMER profile model. There is also a greater overlap between globins and non-globins in the HMMER profile model compared with the fuzzy model. There are 3 globins overlapping with non-globins for Z-scores varying from 1.0 to 2.0 in the HMMER profile model. In contrast, the fuzzy model has no globins in this range. Figures 8 and 9 show the plots of sensitivity and specificity with respect to Z-scores for both classical and fuzzy profile models. The plots demonstrate that the fuzzy model performs better than the classical model.

## Evaluation on kinases

We repeated the evaluation on the kinase family. Kinases are enzymes belonging to a very extensive family of proteins, which share a conserved catalytic core common with both serine/threonine and tyrosine. They are responsible for transferring a phosphate group from a phosphate donor onto an acceptor amino acid in a substrate protein. Kinases have been extensively studied by Taylor (*23*) and Krogh (*19*). Complete protein kinase catalytic domains range from 250 to 300 residues. The kinase dataset used in this study consists of 126 sequences with 72 representative sequences. A total of 1,141 non-kinase sequences extracted from SWISS-PROT database (*20*) were included in the test dataset. The kinase sequences range in length from 100 to 800 amino acids and the profile model was built using 5-fold cross validation. From the Z-score plots shown in Figures 10 and 11 for classical and fuzzy kinase profile models, respectively, similar trends were also observed in the kinase family as observed in the globin family.

# Conclusion

We have proposed a fuzzy profile HMM based on Choquet integrals and Sugeno fuzzy measures to overcome the limitations of statistical independence in classical HMMs and to achieve an improved alignment and

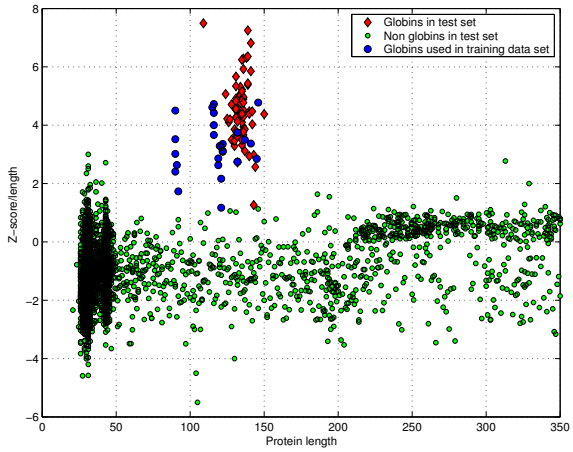


**Fig. 6** Normalized Z-score obtained for globin and non-globin sequences from the classical profile HMM (HMMER) against protein chain length.

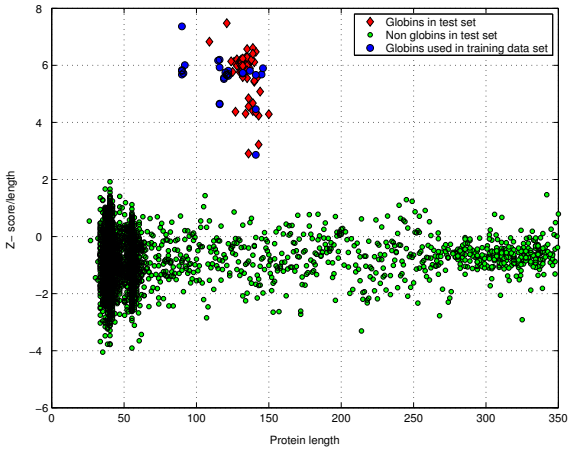


**Fig. 7** Normalized Z-score obtained for globin and non-globin sequences from the fuzzy profile HMM against protein chain length.

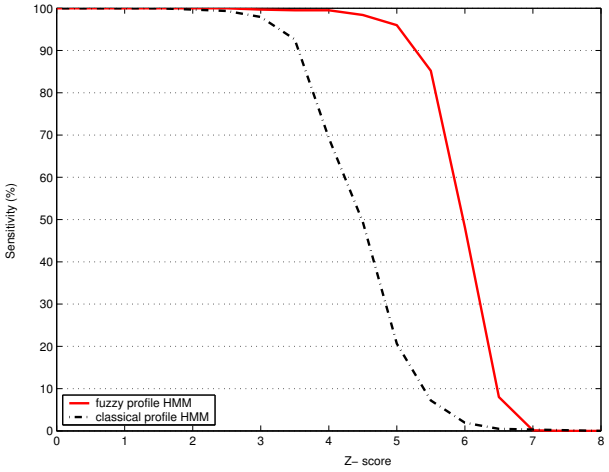


**Fig. 8** Sensitivity of globin and non-globin sequences against Z-score for the fuzzy and classical profile HMMs.

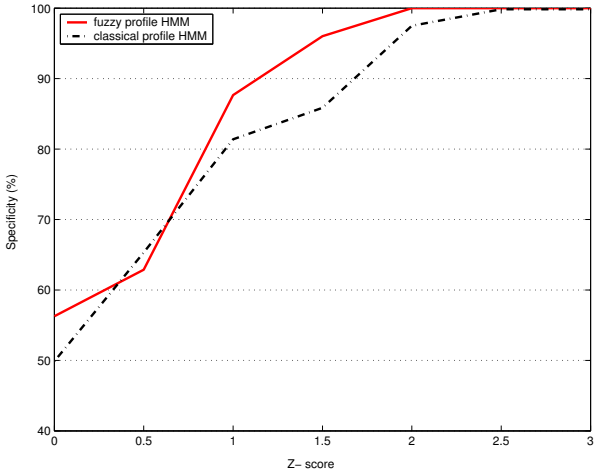


**Fig. 9** Specificity of globin and non-globin sequences against Z-score for the fuzzy and classical profile HMMs.

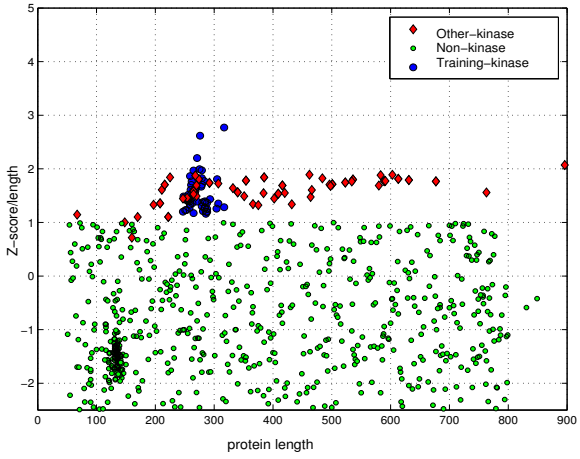


**Fig. 10** Normalized Z-score obtained for kinase and non-kinase sequences from the classical profile HMM (HMMER) against protein chain length.

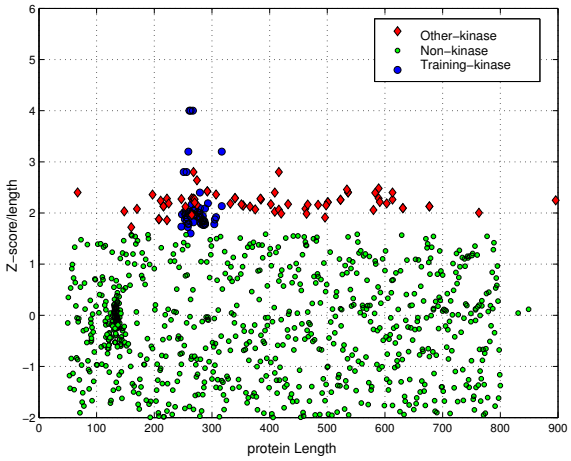


**Fig. 11** Normalized Z-score obtained for kinase and non-kinase sequences from the fuzzy profile HMM against protein chain length.

better log-odd scores for the sequences belonging to a given family. The estimation of Choquet integrals takes into account the ascending values of the scores of the neighboring states while calculating the scores for a given state, hence providing better alignment and improved log-odd scores and Z-scores. The proposed fuzzy profile HMM was tested on the globin and kinase families and compared with the classical profile model. The obtained results establish the superiority of the fuzzy profile HMM. In addition, the fuzzy profile model produces more accurate biologically significant alignments than the classical model because of the relaxation of the statistical independence assumption. Our future study will try to make the fuzzy measures more effective by taking into account the relative importance of biological and physio-chemical factors of the family.

## Authors' contributions

NPB conceived the initial idea of the proposed approach, collected the datasets, conducted experiments and prepared the manuscript. MC and JK refined the idea, supervised the project and revised the manuscripts. All authors read and approved the final manuscript.

## Competing interests

The authors have declared that no competing interests exist.

# References


1. Eddy, S.R. Profile hidden Markov models. 1998. *Bioinformatics* 14: 755-763.
2. Henikoff, S. and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* 89: 10915-10919.
3. Baldi, P. and Brunak, S. 2001. *Bioinformatics: The Machine Learning Approach* (second edition). MIT Press, Cambridge, USA.
4. Churchill, G.A. 1989. Stochastic models for heterogeneous DNA sequences. *Bull. Math. Biol* 51: 79-94.
5. Durbin, R., *et al.* 1998. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, Cambridge, UK.
6. Koski, T. 2001. *Hidden Markov Models for Bioinformatics.* Kluwer Academic Publishers, Dordrecht, the Netherlands.
7. Mohamed, M.A. and Gader, P. 2000. Generalized hidden Markov models. I. Theoretical frameworks. *IEEE Trans. Fuzzy Systems* 8: 67-81.
8. Sugeno, M. 1977. Fuzzy measures and fuzzy integrals: a survey. In *Fuzzy Automata and Decision Processes* (eds. Gupta, M.M., *et al.*), pp. 89-102. North-Holland, New York, USA.
9. Tran, D. and Wagner, M. 1999. Fuzzy hidden Markov models for speech and speaker recognition. In *Proceedings of the 18th International Conference of the North American Fuzzy Information Society*, pp. 426-430. New York, USA.
10. Cheok, A.D., *et al.* 2001. Use of a novel generalized fuzzy hidden Markov model for speech recognition. In *Proceedings of the 10th IEEE International Conference on Fuzzy Systems*, pp. 1207-1210. Melboune, Australia.
11. Shi, H. and Gader, P.D. 1996. Lexicon-driven handwritten word recognition using Choquet fuzzy integral. In *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, Vol. 1, pp. 412-417. Beijing, China.
12. Bidargaddi, N.P., *et al.* 2004. Fuzzy decoding in profile hidden Markov models for protein family identification. In *Advances in Bioinformatics and its Applications*, pp. 119-131. Fort Lauderdale, USA.
13. Bidargaddi, N.P., *et al.* 2005. A fuzzy Viterbi algorithm for improved sequence alignment and searching of proteins. *Lect. Notes Comput. Sci.* 3449: 11-21.
14. Gribskov, M., *et al.* 1987. Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA* 84: 4355-4358.
15. Krogh, A. 1998. An introduction to hidden Markov models for biological sequences. In *Computational Methods in Molecular Biology* (eds. Salzberg, S.L., *et al.*), pp. 45-63. Elsevier Science, Amsterdam, the Netherlands.
16. Wang, Z. and Klir, G.J. 1992. *Fuzzy Measure Theory.* Plenum Press, New York, USA.
17. Grabisch, M., *et al.* 2000. *Fuzzy Measures and Integrals: Theory and Applications.* Physica-Verlag, Berlin, Germany.
18. Bashford, D., *et al.* 1987. Determinants of a protein fold. Unique features of the globin amino acid sequences. *J. Mol. Biol.* 196: 199-216.
19. Krogh, A., *et al.* 1994. Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.* 235: 1501-1531.
20. Boeckmann, B., *et al.* 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 31: 365-370.

21. Sonnhammer, E.L., *et al.* 1997. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* 28: 405-420.
22. Bateman, A. 2002. The Pfam protein families database. *Nucleic Acids Res.* 30: 276-280.
23. Taylor, W.R. 1986. Identification of protein sequence homology by consensus template alignment. *J. Mol. Biol.* 188: 233-258.