BMC Bioinformatics

BioMed Central

Methodology article

Open Access

HMM Logos for visualization of protein families

Benjamin Schuster-Böckler[1,2], Jörg Schultz[3] and Sven Rahmann*[1,4]

Address: [1]Department of Mathematics and Computer Science Freie Universität Berlin, Germany, [2]Present address: Prundsbergstr. 23a, D-82064 Strasslach, Germany, [3]Department of Bioinformatics, Biozentrum, Universität Würzburg, Am Hubland, D-97074 Würzburg, Germany and [4]Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestr. 73, D-14195 Berlin, Germany

Email: Benjamin Schuster-Böckler - bendaboots@gmx.de; Jörg Schultz - Joerg.Schultz@biozentrum.uni-wuerzburg.de; Sven Rahmann* - Sven.Rahmann@molgen.mpg.de

* Corresponding author

Published: 21 January 2004

BMC Bioinformatics 2004, **5**:7

Received: 07 November 2003
Accepted: 21 January 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/7

Abstract

Background: Profile Hidden Markov Models (pHMMs) are a widely used tool for protein family research. Up to now, however, there exists no method to visualize all of their central aspects graphically in an intuitively understandable way.

Results: We present a visualization method that incorporates both emission and transition probabilities of the pHMM, thus extending sequence logos introduced by Schneider and Stephens. For each emitting state of the pHMM, we display a stack of letters. The stack height is determined by the deviation of the position's letter emission frequencies from the background frequencies. The stack width visualizes both the probability of reaching the state (the hitting probability) and the expected number of letters the state emits during a pass through the model (the state's expected contribution).

A web interface offering online creation of HMM Logos and the corresponding source code can be found at the Logos web server of the Max Planck Institute for Molecular Genetics http://logos.molgen.mpg.de.

Conclusions: We demonstrate that HMM Logos can be a useful tool for the biologist: We use them to highlight differences between two homologous subfamilies of GTPases, Rab and Ras, and we show that they are able to indicate structural elements of Ras.

Background

Introduction

Many existing gene or protein sequences in different organisms are related through evolution and can be grouped into families. One way of representing such a family is through a *profile Hidden Markov Model* (pHMM). A pHMM is a fully probabilistic generative model; it specifies position-specific letter emission distributions and also position-specific insertion and deletion probabilities to describe a sequence family. The existence of efficient algorithms for pHMM creation and database search [1] makes pHMMs the tool of choice for protein family research. For example, the protein family and domain databases Pfam [2] and SMART [3] both use pHMMs. However, the large number of parameters in the underlying model makes it non-trivial to present a visual overview of the characteristics that make up a family.

For *sequence profiles*, also known as position-specific score matrices (PSSMs), and ungapped multiple alignments, there exists a visualization method called the *Sequence Logo* [4]. A Sequence Logo graphically represents the conservation of the columns (positions) in a multiple alignment by plotting a stack of letters (nucleotides or amino acids) for each position. The total stack height is computed as the *information content* of the column, i.e., its relative entropy distance from an assumed background distribution. The relative height of each letter in the stack is proportional to its frequency at the position. Usually, colors are used to represent different properties of the letters (e.g., green for aromatic amino acids).

If we ignore the position-specific insertion and deletion probabilities of a pHMM, we can treat is as a PSSM and visualize it with a sequence logo (the makelogo tool of the SAM software package [5] does exactly this), but this would mean throwing away a substantial part of the available information. Therefore our aim is to modify Sequence Logos in such a way that they give an impression of the central aspects of pHMMs: Which positions can be deleted, which ones are highly conserved, and where can we expect long insertions?

Profiles and sequence logos

Let Σ be an alphabet and $|\Sigma|$ its cardinality. For DNA, $|\Sigma| = 4$, and the letters of the alphabet are the four nucleotides A, C, G, and T. For proteins, $|\Sigma| = 20$, and the letters are the twenty amino acids.

A *profile* is a probabilistic description of a sequence. It specifies a probability distribution over the alphabet's letters for each position. More formally, a profile P of length L over Σ is an $L \times |\Sigma|$ matrix (P_{ij}) $(i = 1,...,L; j \in \Sigma)$, such that $P_{ij} \geq 0$ for all i, j and $\Sigma_{j\in\Sigma} P_{ij} = 1$ for all i.

A multiple sequence alignment of N sequences with L columns or positions can be interpreted as a profile. Let C_{ij} be the number of occurrences of letter $j \in \Sigma$ at position i, and let $N_i := \Sigma_{j\in\Sigma} C_{ij} \leq N$ be the number of non-gap letters at position i. Then the maximum likelihood (ML) estimation of the profile P associated with this alignment is given by $P_{ij} := C_{ij}/N_i$. When the multiple alignment contains only few sequences, ML estimation results in many "impossibilities" (zero probabilities) in the profile and hence in over-fitting the model to the small sample. To counteract this problem, the profile is regularized, either by using Dirichlet mixture priors [6], or by alternative techniques (e.g., [7]).

The *uncertainty* or *entropy* [8] of distribution P_i at the i-th position of the profile is given by $H(P_i) = -\Sigma_{j\in\Sigma} P_{ij} \log_2 P_{ij}$. The entropy $H(P_i)$ is always nonnegative. It vanishes if and only if P_i is a Dirac distribution, i.e., if the whole mass is accumulated at a single letter. The entropy takes its maximal value of $\log_2 |\Sigma|$ bits (2 bits for DNA, approximately 4.32 bits for proteins) when P_i is the uniform distribution, i.e., when $P_{ij} = 1/|\Sigma|$ for all j [9]. Since we use the binary logarithm $\log_2$, the unit of the entropy is called a "bit". When we use the natural logarithm, it is called a "nat", and for $\log_{10}$, it is called a "dit".

We may define the *information content* $I(P_i)$ of position i as the "opposite" of its uncertainty,

$$I(P_i) := \log_2 |\Sigma| - H(P_i) = \log_2 |\Sigma| + \sum_{j\in\Sigma} P_{ij} \log_2 P_{ij}. \qquad (1)$$

The information content is a number between 0 and $\log_2 |\Sigma|$ bits and measures the conservation of a position in a profile.

Since conserved positions in sequence families are considered to be functionally or structurally important, they should stand out when the profile is visualized. Schneider and Stephens [4] achieved this goal by representing each position by a stack of letters, where the stack height at position i is precisely the information content $I(P_i)$.

While this method works well on DNA alignments, additional considerations must be made for protein sequences. Amino acids naturally occur with different "background" frequencies. For example, tryptophan (W) occurs much less frequently than leucine (L). The background frequencies might be computed by counting amino acid occurrences in all known proteins, or only in the proteins of the superfamily under consideration. Assume that the background frequency of amino acid j is $\pi_j > 0$. Then the important positions are those whose distribution differs from π. Therefore it has become common practice to consider the *relative entropy* between the distributions P_i and π,

$$H(P_i \,||\, \pi) = \sum_{j\in\Sigma} P_{ij} \cdot \log_2(P_{ij}/\pi_j), \qquad (2)$$

where $0 \cdot \log_2(0/\pi_j) := 0$ by continuity as long as $\pi_j > 0$.

Note that for the uniform distribution $\pi_j = 1/|\Sigma|$, we have $H(P_i \,||\, \pi) = I(P_i)$. Thus the information content of P_i, as defined above, is its relative entropy distance from the uniform distribution.

In a classical Sequence Logo, the stack height at position i is $H(P_i \,||\, \pi)$, the height of letter j within the stack is $P_{ij} H(P_i \,||\, \pi)$, the letters are stacked in sorted order, the largest letter being on top of the stack, and colors may be used to highlight different properties of different letters. The HMM Logo inherits all of these characteristics, but also

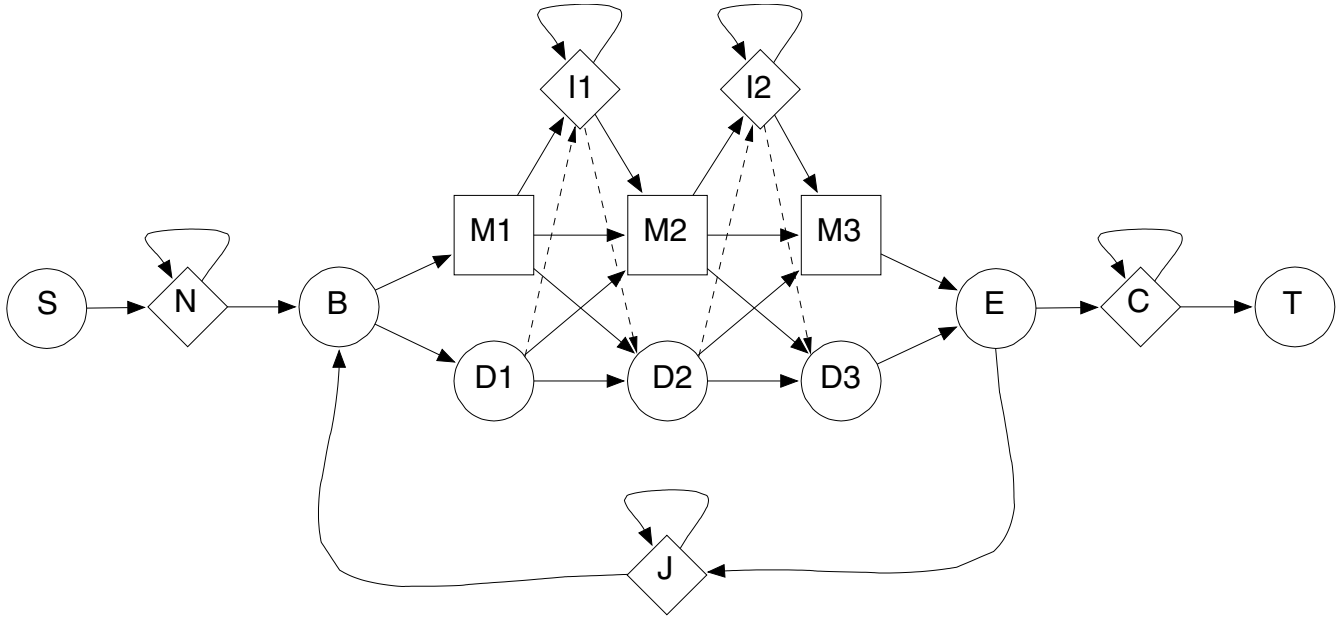

Figure 1
A model of a profile HMM of length 3. Transitions marked by solid arrows constitute the Plan7 model used by HMMER [12]. In the SAM model [5], additionally $D \rightarrow I$ and $I \rightarrow D$ transitions (dashed arrows) are possible.

has additional ones to represent the additional information contained in a pHMM.

Profile HMMs

An HMM is a discrete time Markov chain that emits a letter from the alphabet Σ whenever a state is visited. The central idea is that only the emitted letters can be observed, but that the state sequence is hidden. A comprehensive review on the topic of HMMs can be found in the literature [10]. Profile HMMs are a specialization of HMMs to represent sequence families. Applications to protein modeling were first described in a paper by Krogh and co-workers [11], and are reviewed, for example, by Eddy [1].

Figure 1 shows the transition graph of a pHMM according to the HMMER software package [12]. For each position i (a consensus column of the underlying multiple alignment), a "match" state M_i models the distribution E_i of emitted letters at that position; it corresponds exactly to the profile distribution P_i. An "insert" state I_i allows for insertion of one or more letters between positions i and $i + 1$; their distribution E_i' is individually specified for each insert state. A "delete" state D_i is non-emitting and allows to pass the corresponding match state M_i, resulting in a deletion at the i-th alignment position. The part consisting of the M_i, I_i, and D_i states, flanked by the B and E states, is called the *main model*. There are further special states (*S*, *N*, *J*, *C*, and *T*) in Figure 1, which are not relevant for HMM Logos. As an exception, the background frequencies π of the letters may be learned from the emission probabilities of the *N* or *C* state, which represent flanking sequence.

A path through the main model starts in the silent (non-emitting) begin state *B*, ends in *E*, and follows the legal state transitions. As in every Markov chain, state transitions have probabilities associated with them. We write $A_{s,t}$ for the transition probability $s \rightarrow t$, so we have $A_{s,t} \geq 0$, $\Sigma_t A_{s,t} = 1$ for all s, and $A_{s,t} = 0$ whenever no arrow $s \rightarrow t$ exists. There are exactly seven outgoing transitions (hence the model name Plan7) from every position i (except the last one): $M_i \rightarrow I_i$, $M_i \rightarrow M_{i+1}$, $M_i \rightarrow D_{i+1}$; $D_i \rightarrow M_{i+1}$, $D_i \rightarrow D_{i+1}$; $I_i \rightarrow M_{i+1}$, and the self-loop $I_i \rightarrow I_i$.

There are two major pHMM software packages, HMMER [12] and SAM [5], with small differences between their model topologies. So far we have described the HMMER model. The SAM model allows more state transitions: In addition to the transitions marked by the solid arrows in Figure 1, the transitions marked by dashed arrows, $D_i \rightarrow I_i$ and $I_i \rightarrow D_{i+1}$ are possible.

Results

HMM Logo concepts

The relevant information contained in a pHMM of length L can be summarized as

- letter background frequencies $\pi = (\pi_j)$, $j \in \Sigma$,
- emission probabilities $E = (E_{ij})$ for match states (M_i), $i = 1,...,L$, $j \in \Sigma$,
- emission probabilities $E' = (E'_{ij})$ for insert states (I_i), $i = 1,...,L - 1$, $j \in \Sigma$,
- state transition probabilities $A = (A_{s,t})$.

Sequence Logos can already take care of visualizing the emission probabilities in comparison to the background frequencies. We shall use the remaining dimension of a stack, its *width*, to visualize the transition probabilities in a meaningful way.

Each path $B \to \ldots \to E$ through the main model visits ("hits") certain states and misses others. For example, a path may hit either M_i or D_i, but not both. When a path hits an insert state I_i, several letters may be emitted before it moves on to M_{i+1}. This leads to the following definitions.

Definition 1 (Hitting probability). Let s be a state of the main model. The *hitting probability* $h(s)$ is the probability that a path $B \to \ldots \to E$ following the transition probabilities A, hits s at some point between B and E.

Definition 2 (Contribution). Let s be a match or insert state of the main model. Its *contribution* $C(s)$ to an emitted sequence is a random variable describing the number of emitted letters in s along a path $B \to \ldots \to E$. Further, we define $c(s) := \mathbb{E}[C(s)]$ as the *expected contribution* of state s.

Computation of hitting probabilities

The hitting probability of a state equals the sum of probabilities of all paths $B \to \ldots \to E$ visiting this state. Because of the self loops in insert states, this is an infinite number of paths. The hitting probability can nevertheless be computed efficiently using a forward-type dynamic programming algorithm as follows.

Proposition 1. Define $\mu_i := A_{I_{i-1},M_i}/(1 - A_{I_{i-1},I_{i-1}})$ as the conditional probability that a path hitting I_{i-1} exits into M_i. Then 1 - μ_i is the probability of exiting into D_i. For the Plan7 model disallowing the $I_{i-1} \to D_i$ transition we have $\mu_i = 1$. For the general SAM-type pHMM model allowing all 9 transitions, the hitting probabilities are

- at the first position given by

$$\begin{aligned} h(M_1) &= A_{B,M_1}, \\ h(D_1) &= A_{B,D_1} &&= 1 - h(M_1), \\ h(I_1) &= h(M_1)\cdot A_{M_1,I_1} + h(D_1)\cdot A_{D_1,I_1} &&= h(M_1)\cdot A_{M_1,I_1} + (1 - h(M_1))\cdot A_{D_1,I_1}. \end{aligned}$$

- at the following positions $i \geq 2$ given by

$$\begin{aligned} h(M_i) &= h(M_{i-1})\cdot A_{M_{i-1},M_i} + h(D_{i-1})\cdot A_{D_{i-1},M_i} + h(I_{i-1})\cdot \mu_i \\ &= h(M_{i-1})\cdot [A_{M_{i-1},M_i} + A_{M_{i-1},I_{i-1}}\cdot \mu_i] \\ &\quad + (1 - h(M_{i-1}))\cdot [A_{D_{i-1},M_i} + A_{D_{i-1},I_{i-1}}\cdot (1 - \mu_i)], \\ h(D_i) &= 1 - h(M_i), \\ h(I_i) &= h(M_i)\cdot A_{M_i,I_i} + h(D_i)\cdot A_{D_i,I_i} = h(M_i)\cdot A_{M_i,I_i} + (1 - h(M_i))\cdot A_{D_i,I_i}. \end{aligned}$$

Proof. The initializations for $h(M_1)$ and $h(I_1)$ are obvious from Figure 1. At every position $i \geq 1$ we have $h(D_i) = 1 - h(M_i)$ because each path passes either through M_{i-1} or D_{i-1}.

For $h(M_i)$, $i \geq 2$, there are three ways into M_i. The first term accounts for paths that come directly from M_{i-1}, the second term similarly accounts for direct entries from D_{i-1}, and the last term accounts for paths that enter via I_{i-1}. A similar argument applies to the insert state hitting probabilities, for which there are only two ways of entry. All probabilities can be expressed solely in terms of $h(I_{i-1})$ as shown. □

Computation of expected contributions

The expected contribution of each state is easily derived from its hitting probability. Since delete states are non-emitting, their contribution is zero.

Proposition 2 (Expected contribution). We have

- $c(M_i) = h(M_i)$,
- $c(I_i) = h(I_i)/(1 - A_{I_i,I_i})$.

Proof. If a match state M_i is hit, it contributes $C = 1$ letter; otherwise, it contributes nothing. This results in an expectation of $c(M_i) = 1 \cdot h(M_i) + 0 \cdot (1 - h(M_i)) = h(M_i)$.

If an insert state I_i is hit, its contribution has a geometric distribution with "success parameter" (probability of leaving the state) $1 - A_{I_i,I_i}$. Then the expected sojourn time is the reciprocal of this probability. If the state is not hit, its contribution is zero. Together, this results in an expectation of $c(I_i) = h(I_i)/(1 - A_{I_i,I_i})$. □

Proposition 3 (Expected number of emitted letters). The expected number of emitted letters during a walk from B to E through a profile HMM with L positions is

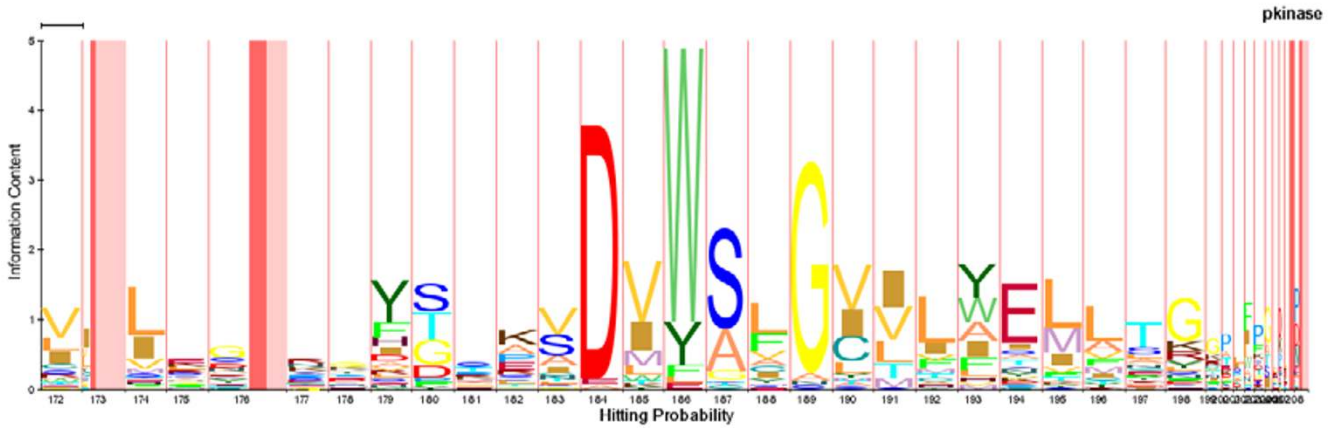

Figure 2
Partial logo (positions 172–209) of the Pfam pkinase model. Positions with narrow match state stacks are likely to be deleted in typical family members. The total width of a red-shaded (dark+light) stack visualizes the expected number of inserted letters. The left dark-shaded part of the stack's width represents the probability that at least one letter is inserted. The difference is illustrated by comparing I_{173} with I_{176}: Both states have approximately the same expected contribution, but the hitting probability of I_{176} is higher. The insertion stack height is zero for all shown examples because the emission probabilities correspond to the background frequencies.

$$\mathbb{E}[\text{Emitted letters}] = \sum_{i=1}^{L-1} [c(M_i) + c(I_i)] + c(M_L).$$

Proof. Every emitting state s emits on average $c(s)$ letters during a walk from state B to E. The proposition follows from the linearity of expectation. □

We find it logical to set the width of the stack of an emitting state s to $c(s)$, for then by Proposition 3, the total width of the logo represents the total number of emitted letters. However, we would like to display both the hitting probability and the expected contribution of a state. This is a non-problem for match states. For insert states s, we always have $h(s) \leq c(s)$, and so we can use two different background shadings for $h(s)$ and for the remainder $c(s)$ - $h(s)$.

HMM Logo layout

The final definition of an HMM logo is as follows; see Figure 2 for a typical example.

• HMM Logos consist of alternating stacks for match and insert states for all positions 1,...,L in the profile; the stack order is M_1, I_1, M_2, I_2,...,I_{L-1}, M_L.

• The total height of a stack is the relative entropy $H(e||\pi)$ between the state's emission distribution e and the background distribution π obtained from state N.

• The relative height of letter $j \in \Sigma$ within the stack is proportional to its emission probability e_j.

• The letters are stacked in sorted order, the largest letter being on top of the stack.

• The total width of a stack s is its expected contribution $c(s)$.

• The background of an insert state's stack is shaded in two different colors for a total width of $c(s)$ "letter units". The first $h(s)$ units represent the hitting probability and are shaded with a medium-red background. The remaining $c(s)$ - $h(s)$ units are shaded with a lighter red.

• The upper left corner of the logo shows a horizontal bar representing a contribution of 1 letter.

• Insert state stacks are always displayed with a width of at least one pixel, thus making consecutive positions easier to distinguish.

• Letters are drawn in different colors. The color scheme depends on the alphabet; amino acids are colored to represent structural or functional similarity.

• The position number is displayed on the x-axis below every match/insert pair. The height of the y-axis is min(2, $max_i \{H(E_i||\pi), H(E_i'||\pi)\}$) bits, i.e., at least 2 bits, even if all stacks have a lower height.

Visualization of subfamily-specific sites

Since profile HMMs are predominantly used for protein family and domain modeling, we present examples that illustrate the utility of HMM Logos in this area.

While building an HMM for a domain, one usually tries to cover all homologous sequences. But, with ongoing experimental characterization, it often becomes clear that a single domain family consists of multiple, functionally divergent subfamilies.

Identifying these subfamilies and characterizing their determinants is an important step in protein function prediction. Creation of subfamily-specific profile HMMs is a first step in this direction performed by domain databases like SMART [3]. HMM Logos highlight regions which distinguish homologous subfamilies from each other and thereby facilitate the detection of subfamily determinants. Here, we illustrate this application on two subfamilies, Ras and Rab, of the small GTPases, whose profile HMMs were obtained from the SMART multiple subfamily alignment.

Combining sequence and structure analysis, Pereira-Leal and Seabra identified five regions which distinguish the Rab proteins from Ras like members [13]. Figure 3 depicts four of these regions (RabF2 to RabF5), which, in the three-dimensional structure, cluster between sheets β3 and β4. These are included in the switch II region, which changes conformation upon binding of GTP or GDP and mediates interactions with effectors and regulators. Furthermore, this region allows interactors to distinguish between Ras and Rab proteins and thus should contain subfamily determinants. By comparing the HMM logos for the two subfamilies, indeed both, domain and subfamily specific sites become apparent. For example, N-terminal to the small GTPase typical sequence DTAG, there is a highly conserved W in the Rab subfamily, whereas the corresponding site in Ras protein shows less conservation but a prevalence of the hydrophobic amino acids L or V.

Highlighting of loop regions

An important feature distinguishing HMM Logos from standard Sequence Logos is their ability to visualize regions with long expected inserts. These insertions usually do not happen within conserved structural elements, that is alpha helices or beta sheets, as this would influence and possibly break the structure of the whole domain. Instead, insertions are more likely to occur within loop regions.

Therefore the presence of frequent insertions at a given site can indicate that the site itself and its neighbors lie within a loop region. Figure 4 illustrates that this concept holds true for the HMM of the Ras domain. Here two regions with a prominent insert state can be found. Mapping them onto the known secondary structure (Protein Data Bank identifier PDB:121P) shows that these insert states indeed fall between the known structural elements.

Discussion

The examples in the previous section illustrate the potential utility of HMM Logos, but they also point out a particularity of the HMMER software: In all Pfam and SMART pHMMs we looked at, the stack height in all insertion states is zero. This seems to be a consequence of HMMER's hmmbuild program: Insert states receive a very high emission prior that is equal to the background. This makes sense to allow the insertion of variable sequence parts of varying lengths at a position, i.e., in an insert state with high expected contribution. In order to change the emission probabilities away from the background, one would have to observe a consistent insertion that is common to several family members at the same position. Then however, hmmbuild would model this conserved "insertion" as a match state and model the sequences skipping this position via the delete state, even if this is the majority of the family members. This is immediately obvious from the numerous narrow match states shown in Figure 2. In our opinion, these narrow match states could be modeled more meaningfully as insert states with non-trivial emission probabilities. So while HMMER supports insert-specific emission probabilities, they do not seem to be used. HMM Logos immediately made this particularity visible; we were not aware of it before.

While we hope that HMM Logos can help to compare families visually, the RAS-RAB example (Figure 3) leaves us asking for more functionality: It would be useful to align two or several logos. In this way, a multiple family alignment of many sequences from a few different subfamilies could be represented as a multiple alignment of a few logos. Finding the most natural definition of the alignment score and the graphical representation of such a Logo alignment seem to be interesting topics for the future.

Conclusion

We have developed a method to visualize profile HMM specific information and demonstrated its utility for the biologist who wants to *look* at the model of a protein family or domain.

A PERL package for parsing and visualizing HMMER pHMMs is available under the GNU General Public License from the authors and can be downloaded from the Logos server of the Max Planck Institute for Molecular Genetics http://logos.molgen.mpg.de. At the same location we also offer the WWW-based tool LogoMat-M for HMM Logo generation which can be accessed in several

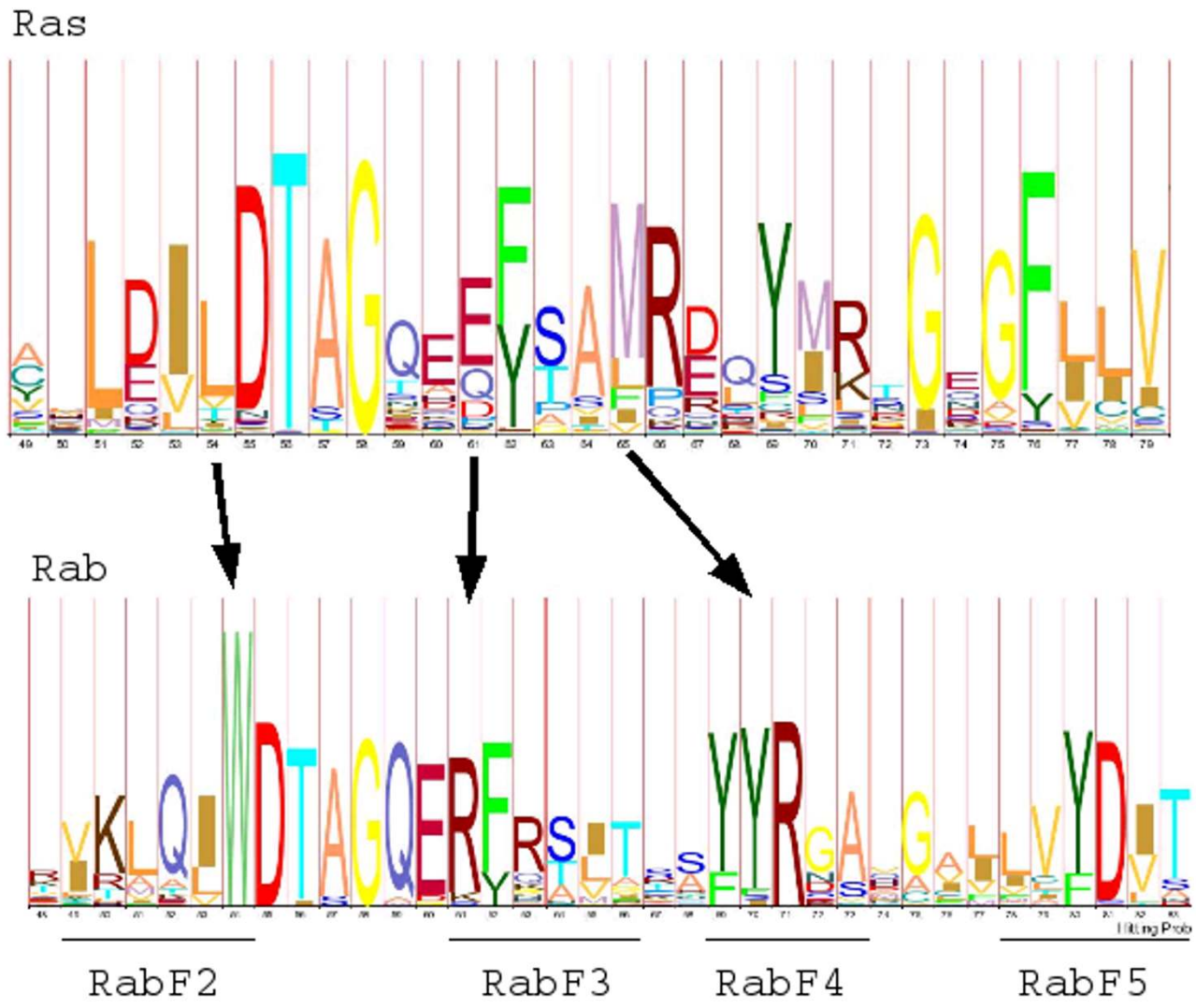

Figure 3
Comparison of the HMM Logos of the small GTPases Ras and Rab from SMART [3]. The Ras logo is based on an alignment of 35 sequences; the Rab logo on 48 sequences. The height of the entire vertical axis is 5 bits for both logos. Subfamily specific sites RabF2 to RabF5 [13] are indicated by arrows.

ways. For example, in an interactive web form the user may specify a file in HMMER format which is uploaded and processed. Available options are described in the online help. Alternatively, it is possible to URL-encode Pfam identifiers, such as in

http://logos.molgen.mpg.de/cgi-bin/logomat-m.cgi?pfamid=AAA.

This will display a logo of the Pfam entry "ATPase family associated with various cellular activities" (AAA), using the default settings. Finally, the logos can be directly accessed from the Pfam website by pressing the "View HMM Logo" button on each domain's or family's overview page.

Authors' contributions

SR had the initial idea to use the stack width to visualize the insertion and deletion probabilities. BSB implemented the software and the web server and invented the two-colored scheme for visualizing both hitting probability and expected contribution of an insert state. This work

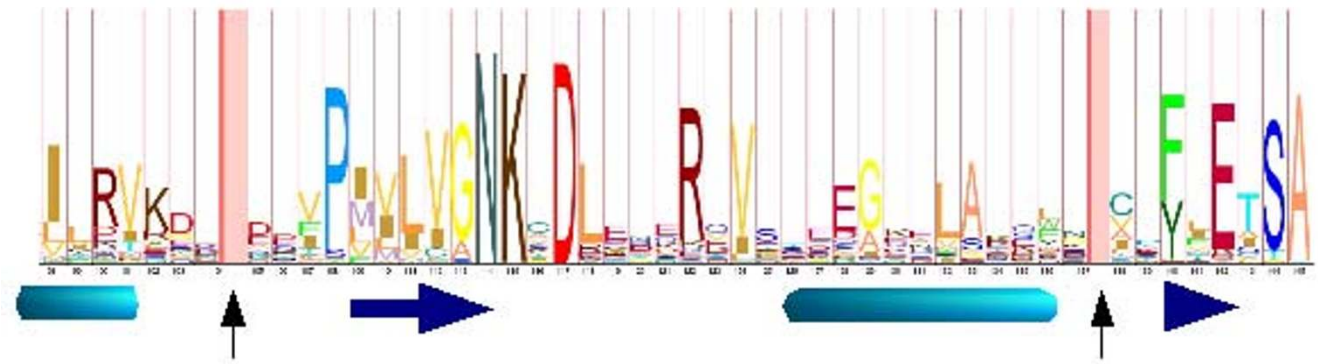

Figure 4
Mapping of structural elements to a region of the Ras family HMM Logo. The mapping was obtained by aligning the sequence of p21 ras, the structure of which has been solved, to the Ras family pHMM. Below the logo, insert regions are highlighted by vertical arrows, and the secondary structure of p21 ras is indicated (alpha helices: barrels; beta sheets: horizontal arrows).

is part of his Bachelor's degree at the Free University of Berlin. JS examined the small GTPases with HMM Logos. All authors read and approved the final manuscript.

Acknowledgements

We would like to express our gratitude to the PERL community; in particular to the creators of the PERL Data Language (PDL) and to the authors of the modules HMMERViewer (Robin Dowell), Imager (Arnar M. Hrafnkelsson and Tony Cook), and TFBS (Boris Lenhard). We thank Andrea Weiβe, Niels Köhler and Victoria Ornelas for valuable comments and discussions, David Studholme from the Sanger Center for information about direct Pfam access, Wilhelm Rüsing for help with the web server, and Martin Vingron for supervising the thesis of BSB. One of the anonymous referees provided valuable comments about applying the method to SAM-style HMMs.

References

1. Eddy SR: **Profile Hidden Markov Models.** *Bioinformatics* 1998, **14:**755-63.
2. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL: **The Pfam protein families database.** *Nucleic Acids Res* 2002, **30:**276-280.
3. Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, Ciccarelli F, Copley RR, Ponting CP, Bork P: **Recent improvements to the SMART domain-based sequence annotation resource.** *Nucleic Acids Res* 2002, **30:**242-244.
4. Schneider TD, Stephens RM: **Sequence Logos: A new way to display consensus sequences.** *Nucleic Acids Res* 1990, **18:**6097-6100.
5. Hughey R, Karplus K, Krogh A: **SAM Sequence Alignment and Modeling Software System. Technical Report UCSC-CRL-99-11, updated for SAM Version 3.4.** *Baskin School of Engineering University of California Santa Cruz*; 2003.
6. Sjølander K, Karplus K, Brown M, Hughey R, Krogh A, Mian IS, Haussler D: **Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology.** *CABIOS* 1996, **12(4):**327-345.
7. Rahmann S, Müller T, Vingron M: **On the Power of Profiles for Transcription Factor Binding Site Detection.** *Stat Appl Genet Mol Biol* 2003, **2(1**Article 7 [http://www.bepress.com/sagmb/vol2/iss1/art7].
8. Shannon CE, Weaver W: **The Mathematical Theory of Communication.** *Urbana: University of Illinois Press* 1949.
9. Cover TM, Thomas JA: **Elements of Information Theory.** *Wiley Series in Telecommunications John Wiley & Sons*; 1991.
10. Rabiner LR: **A tutorial on Hidden Markov Models and selected applications in speech recognition.** *Proc IEEE* 1989, **77:**257-286.
11. Krogh A, Brown M, Mian IS, Sjølander K, Haussler D: **Hidden markov models in computational biology. Applications to protein modeling.** *J Mol Biol* 1994, **235:**1501-31.
12. Eddy SR: **HMMER User's Guide: Biological sequence analysis using profile Hidden Markov Models, version 2.2.** *Washington University School of Medicine* 2001 [http://hmmer.wustl.edu].
13. Pereira-Leal JB, Seabra MC: **The mammalian Rab family of small GTPases: definition of family and subfamily sequence motifs suggests a mechanism for functional specificity in the Ras superfamily.** *J Mol Biol* 2000, **301:**1077-1087.