The following is supplementary material, to be read after the manuscript

M.G. Kann, S.L. Sheetlin, Y. Park, S.H. Bryant, and J.L. Spouge
"Accurate Semi-Global Alignment Statistics and the Improved Identification of Protein Domains"

has been read. It shows that the GLOBAL alignment graph approximates a particularly simple Hidden Markov Model (HMM), and that the GLOBAL algorithm is the Viterbi algorithm for the same HMM.

Caption for the Figure: A diagram of the $a$-th block in the “GLOBAL HMM”

The dotted rectangle bounds an HMM sub-unit equivalent to the $a$-th block in GLOBAL. Circles represent hidden Markov states, which emit symbols (not shown). Arrows indicate possible transitions between the states. Probabilities for emission and transition are not shown.

Text

The following: (1) describes the HMM transitions modeling the $a$-th block; (2) describes the HMM transitions modeling an entire protein sequence; (3) constructs HMM transition probabilities corresponding to the 0 gap penalties in GLOBAL; and (4)
demonstrates that the GLOBAL algorithm for finding the path of maximum log-odds weight is essentially the Viterbi algorithm for finding the HMM path and emissions of maximum probability producing a protein sequence. Beyond the importance of providing GLOBAL with a sound statistical foundation in the theory of HMMs, the construction below suggests that the GLOBAL approximation can provide accurate p-values for some other specialized HMMs.

To describe the HMM transitions modeling the $a$-th block, the first HMM state corresponding to the $a$-th block is $\text{BEG-}\ a$, on the left in the diagram. The next state can be any $\text{COL-}\ i$ ($i = 1,...,m_a$), $\text{GAP-}\ a$, or $\text{BEG-}\ (a+1)$. From each $\text{COL-}\ i$ ($i = 1,...,m_a - 1$), the next state can be $\text{COL-}\ (i+1)$, $\text{GAP-}\ a$, or $\text{BEG-}\ (a+1)$. From $\text{COL-}\ m_a$, the next state can be $\text{GAP-}\ a$ or $\text{BEG-}\ (a+1)$. From $\text{GAP-}\ a$, the next state can be $\text{BEG-}\ (a+1)$ state or a re-entry to $\text{GAP-}\ a$ through a “loop transition”.

To describe the HMM transitions modeling an entire protein sequence, the first HMM state corresponding to a sequence starts in an initial “$\text{BEG-}\ 0$” state leading solely to a “$\text{GAP-}\ 0$” state (with outputs like $\text{GAP-}\ a$ in the diagram, but without inputs). The two states $\text{BEG-}\ 0$ and $\text{GAP-}\ 0$ are therefore equivalent to a block without $\text{COL-}\ i$ states. Except for its loop transition, $\text{GAP-}\ 0$ leads solely to the $\text{BEG-}\ 1$ state, corresponding to the first CDD block. $\text{BEG-}\ 1$ is the start of a series of sub-units like the one diagrammed, each sub-unit corresponding to a CDD block ($a = 1,...,b$). The $b$-th unit terminates with a transition into a “$\text{BEG-}\ (b+1)$” state (with inputs like $\text{BEG-}\ (a+1)$ in the diagram, but without outputs), which also terminates the sequence.
To construct a class of HMM transition probabilities corresponding to the 0 gap penalties in GLOBAL, the following notation is pertinent. Fix the block index $a$, and denote the transition probability from state $x$ to state $y$ within the $a$-th HMM sub-unit $(a=1,...,b)$ by $p_{x,y}$. Let the subscript 0 correspond to BEG-$a$; $i=1,...,m_a$, to COL-$i$; $g$, to GAP-$a$; and $t$, to the terminus BEG-$(a+1)$. The transition probability from BEG-$a$ to COL-$i$ is therefore $p_{0,i}$; from GAP-$a$ to BEG-$(a+1)$, $p_{g,t}$; etc. Intuitively, the 0 gap penalties in GLOBAL correspond to “path-invariance” in the HMM, where every HMM path has about the same probability, if emissions are ignored. Under such conditions, the transition probabilities are mostly irrelevant to the Viterbi algorithm, which then finds the HMM path maximizing only the sequence emission probability.

The loop transition complicates the construction of a path-invariant HMM. To render the loop transition irrelevant, impose the condition that the transition probability $p_{g,t} = \varepsilon$ from GAP-$a$ to BEG-$(a+1)$ be small, making the loop transition almost certain: $p_{g,g} = 1 - \varepsilon$. If $\varepsilon > 0$ is sufficiently small, the loop transition makes a negligible contribution to the probability of any specific path. Accordingly, consider the HMM diagram without the loop transition.

Path-invariance implies as follows that the states have “exponential potentials” $\nu_0 > \nu_i > \ldots > \nu_m > \nu_g > \nu_t > 0$ associated with them. Except for the loop transition, if an edge joins states $x$ and $y$, let $p_{x,y} = \nu_y / \nu_x$, e.g., $\nu_t / \nu_g = \varepsilon$. The state $x$ has “potential” $\log \nu_x$ (drawing from the meaning in physics), so all paths starting at $x$ and ending at $y$
yield the same product $v_y/v_x$ of transition probabilities. The relative magnitude of the potentials $\{v_x\}$ enforces $0 < p_{x,y} < 1$.

Probability is conserved going out of every vertex, so $\sum_{i=1}^{m_a} p_{0,i} + p_{0,g} + p_{0,t} = 1$, $p_{i+1,i} + p_{i,g} + p_{i,t} = 1$ $(i = 1, \ldots, m_a - 1)$, and $p_{m_a,g} + p_{m_a,t} = 1$. Thus, $\sum_{i=1}^{m_a} v_i + v_g + v_t = v_0$, $v_{i+1} + v_g + v_t = v_i$ $(i = 1, \ldots, m_a - 1)$, and $v_g + v_t = v_{m_a}$. The first constraint becomes

$$\left\{ \frac{1}{2} m_a \left( m_a + 1 \right) + 1 \right\} (v_g + v_t) = v_0,$$

so the unique solution for the path-invariant transition probabilities in the GLOBAL HMM are

$$p_{0,i} = v_i / v_0 = \left( m_a - i + 1 \right) \left[ \frac{1}{2} m_a \left( m_a + 1 \right) + 1 \right]^{-1}$$

$$p_{0,g} = v_g / v_0 = \left[ \left( 1 + \varepsilon \right) \left( \frac{1}{2} m_a \left( m_a + 1 \right) + 1 \right) \right]^{-1}$$

$$p_{0,t} = v_t / v_0 = \varepsilon \left( \left( 1 + \varepsilon \right) \left( \frac{1}{2} m_a \left( m_a + 1 \right) + 1 \right) \right)^{-1}$$

$$p_{i+1,i} = v_{i+1} / v_i = \left( m_a - i \right) / \left( m_a - i + 1 \right)$$

$$p_{i,g} = v_g / v_i = \left( 1 + \varepsilon \right)^{-1} / \left( m_a - i + 1 \right)$$

$$p_{i,t} = v_t / v_i = \varepsilon \left( 1 + \varepsilon \right)^{-1} / \left( m_a - i + 1 \right)$$

$$p_{g,t} = v_t / v_g = \varepsilon$$

Thus, the probability of any particular path from BEG-$a$ to BEG-$a+1$ not containing the GAP-$a$ loop transition is a constant, $p_{0,a} = \varepsilon \left[ \left( 1 + \varepsilon \right) \left( \frac{1}{2} m_a \left( m_a + 1 \right) + 1 \right) \right]^{-1}$. Because loop transitions have little effect on path probabilities, the Viterbi algorithm then finds HMM paths maximizing mainly the sequence emission probability.

To bring the diagram’s emission probabilities into the discourse, the BEG-$a$ and BEG-$a+1$ states are silent and do not emit a character. The COL-$i$ state emits a letter following the probabilities for the corresponding column in the CDD. The gap state emits
letters following the background probability, namely, the Robinson-Robinson frequencies.

The correspondence between the HMM states and vertices on the path $\pi$ in Figure 2 can now be given. The initial BEG-0 state corresponds to $(0,0)$. The next transition is from BEG-0 to GAP-0. For $a = 1, \ldots, b$, successive loop transitions from GAP-$(a-1)$ into GAP-$(a-1)$ correspond to the northward sub-path of $\pi$ immediately preceding the $a$-th block, which is terminated by a transition into BEG-$a$. A transition into some COL-i corresponds to the initial eastward sub-path of $\pi$ within the $a$-th block. Successive transitions into successive COL-i states correspond to the northeastward sub-path of $\pi$ within the $a$-th block. The transition into GAP-$a$ corresponds to the final eastward sub-path of $\pi$ within the $a$-th block. Successive transitions into GAP-$(b+1)$ state correspond to the northward sub-path of $\pi$ following the $b$-th block $(a = 1, \ldots, b)$, which is terminated by a transition into BEG-$(b+1)$, corresponding to the vertex $(m,n)$.

To establish the correspondence between the Viterbi and GLOBAL algorithms, note that the emitted sequence is constant during the Viterbi algorithm. Divide the probability of an HMM path and its emissions, therefore, by the (constant) background probability of the emitted sequence, namely the product of its letters’ Robinson-Robinson probabilities. After subtracting and then taking logarithms, the Viterbi algorithm approximately finds the path and emissions maximizing the sum of the log-odds emission scores from the GLOBAL HMM. The GAP-$a$ state emits letters with Robinson-Robinson frequencies, contributing an irrelevant log-odds emission score of 0, and BEG-$a$ is silent. Thus,
Viterbi algorithm maximizes only the sum of the log-odds emission scores from the COL-\(i\) states, so it maximizes essentially the same quantity as the GLOBAL algorithm.

To model a CD, HMMer fits many more HMM parameters than GLOBAL. Although both fit emission probabilities for states equivalent to COL-\(i\) and GAP-\(a\) (probably with similar results), GLOBAL determines its HMM transition probabilities uniquely from Eq (1). In contrast, HMMer fits its transition probabilities (an enormous number) from data. Despite HMMer’s overwhelmingly more flexible fit, however, GLOBAL matches HMMer’s CDD retrieval performance. Thus, HMM emission probabilities likely contribute more to CDD retrieval performance than HMM transition probabilities, if the hidden states have transitions compatible with the data. To support this contention further, note that CDD blocks (by definition) are conserved. Perfect conservation of blocks yields \(p_{0,1} = p_{1,2} = \ldots = p_{m_a-1,m_a} = 1\), not notably in agreement with Eq (1). Despite the disagreement between the transition probabilities for GLOBAL and for perfect block conservation, GLOBAL still has good CDD retrieval performance.

Finally, the GLOBAL HMM is likely to emit long gaps (above, \(p_{g,\varepsilon} = 1 - \varepsilon\) for small \(\varepsilon > 0\)). More than in protein applications, therefore, it might provide a realistic probability model in searches for “genomic signatures” (e.g., of pathogenic viruses), i.e., in situations where gapless blocks of conserved DNA are separated by indefinitely long stretches of the genomic sequence.