BIOINFORMATICS

Vol. 00 no. 00 2007
Pages 1–4

De novo protein repeat identification by probabilistic consistency: Supplementary material

A. Biegert,[1,2] and J. Söding[1,2*]

[1] Department for Protein Evolution, Max Planck Institute for Developmental Biology, Spemannstr. 35, 72076 Tübingen, Germany; [2] Gene Center Munich, University of Munich (LMU), Feodor-Lynen-Str. 25, 81377 Munich, Germany

1 POSTERIOR PROBABILITIES FOR LOCAL HMM-HMM COMPARISON

To quantify the local reliability of an alignment, we would like to calculate posterior probabilities for local HMM-HMM comparison. Assume we align two HMMs q and p of length L_q, L_p. We adopt the thermodynamic interpretation of HMM-HMM alignment that was developed for sequence-sequence alignment by Miyazawa (1995); Kschischo and Lssig (2000); Mückstein *et al.* (2002). The probability for an alignment $\mathcal{A}$ between q and p is given by

$$P(\mathcal{A}) = \frac{e^{\beta S(\mathcal{A})}}{Z} \,, \tag{1}$$

where $S(\mathcal{A})$ is the score for alignment $\mathcal{A}$ and $\beta = 1/kT$ (T = temperature, k = Boltzmann constant) can be assumed to be 1 in the following. Z is the *partition function*:

$$Z = \sum_{\mathcal{A}} e^{\beta S(\mathcal{A})}. \tag{2}$$

The sum runs over all aligments including the empty alignment. Z is effectively a normalization constant in eq. (1). The score $S(\mathcal{A})$ is defined as in (Söding, 2005), eq. (2):

$$S(\mathcal{A}) = \log \sum_{x_1,\ldots,x_L} \frac{P(x_1,\ldots,x_L|\mathrm{co-emitted on}\mathcal{A})}{P(x_1,\ldots,x_L|\mathrm{Random})} \,. \tag{3}$$

The sum runs over all sequences $x_1,\ldots,x_L$ co-emitted along the alignment $\mathcal{A}$ of q and p. This score is a hybrid between the Forward score, which is similar but contains a sum over all alignments within the logarithm, and the Viterby score *sensu strictu*, which would contain a maximum over all co-emitted sequences instead of a sum over them. It is intuitively clear that the summation over $x_1,\ldots,x_L$ improves the score while still allowing to calculate the score with the faster Viterbi algorithm instead of the slower Forward-Backward algorithm. In (Söding, 2005), eq. (3), it was shown that this score simplifies to

$$S(\mathcal{A}) = \log \sum_{l=1}^{L_{MM}} \log\left(\rho_{aa}(q_{i_{\mathcal{A}}(l)}, p_{j_{\mathcal{A}}(l)}\right) \;+\; P_{tr}(\mathcal{A}) \,. \tag{4}$$

*to whom correspondence should be addressed

The sum runs over all L_{MM} pairs of aligned Match-Match states $(M^q_{i_{\mathcal{A}}(l)}, M^p_{j_{\mathcal{A}}(l)})$. We have used the column score

$$\log\left(\rho_{aa}(q_i, p_j)\right) = \log \sum_{a=1}^{20} \frac{q_i(a), p_j(a)}{f(a)} \tag{5}$$

that measures the similarity of match state i of q and match state j of p. It is a weighted version of co-emission probability, with weights being equal to $1/f(a)$. The background frequencies $f(a)$ have entered through the random sequence model probability for the co-emmited sequence, $P(x_1,\ldots,x_L|\mathrm{Random}) = \prod_l f(x_l)$. Their effect is to weigh up rare amino acids in the column score. This makes sense since rare amino acids are less likely to be co-emitted by chance. The product of all transition probabilities along the alignment $\mathcal{A}$ is abreviated by $P_{tr}(\mathcal{A})$.

We now define the Forward and Backward partition functions in complete analogy to the Forward and Backward probabilities for the case of global HMM-sequence alignment (Durbin *et al.*, 1998):

$$F_{MM}(i,j) = \sum_{\mathcal{A}\in\mathcal{F}_{i,j}} e^{\beta S(\mathcal{A})} \tag{6}$$

and similiarly

$$B_{MM}(i,j) = \sum_{\mathcal{A}\in\mathcal{B}_{i,j}} e^{\beta S(\mathcal{A})}. \tag{7}$$

Here, $\mathcal{F}_{i,j}$ is the set of all local alignments between $q_{1\ldots i}$ and $p_{1\ldots j}$ ending in the aligned pair state (M^q_i, M^p_j). Similarly, $\mathcal{B}_{i,j}$ is the set of all local alignments between $q_{i+1\ldots L_q}$ and $p_{j+1\ldots L_p}$ starting *after* the aligned pair state (M^q_i, M^p_j). From these definitions and eq. (1), the (posterior) probability for pair state (M^q_i, M^p_j) to be part of an alignment between q and p is

$$P(M^q_i \diamond M^p_j) = \frac{F_{MM}(i,j) \times B_{MM}(i,j)}{Z} \,. \tag{8}$$

Furthermore, Z can be expressed in terms of the Forward partition function

$$Z = 1 + \sum_{i,j} F_{MM}(i,j) \,, \tag{9}$$

where the 1 is contributed by the empty alignment.

To compute the Forward partition function F_{MM}, we need five dynamic programming matrices F_{XY}, one for each pair state $XY \in$

$\{MM, MI, IM, DG, GD\}$ (see Söding (2005) for details about HMM pair states). We begin the alorithm by initializing the top row and left column of the $F_{MM}(i,j)$ matrix to 1 and proceed to fill the matrix recursively from top left to bottom right using the following recursion relations:

$$\begin{aligned} F_{MM}(i,j) = \rho_{aa}(q_i,p_j) \;\times\; & \\ (\quad & 1 \\ + \quad & F_{MM}(i-1,j-1)\; q_{i-1}(M,M)\; p_{j-1}(M,M) \\ + \quad & F_{MI}(i-1,j-1)\; q_{i-1}(M,M)\; p_{j-1}(I,M) \\ + \quad & F_{IM}(i-1,j-1)\; q_{i-1}(I,M)\; p_{j-1}(M,M) \\ + \quad & F_{DG}(i-1,j-1)\; q_{i-1}(D,M)\; p_{j-1}(M,M) \\ + \quad & F_{GD}(i-1,j-1)\; q_{i-1}(M,M)\; p_{j-1}(D,M) \quad) \end{aligned} \tag{10}$$

$$\begin{aligned} F_{GD}(i,j) = {} & F_{MM}(i,j-1)\; q_{j-1}(M,D) + \\ & F_{DG}(i,j-1)\; q_{j-1}(D,D) \end{aligned} \tag{11}$$

$$\begin{aligned} F_{IM}(i,j) = {} & F_{MM}(i,j-1)\; q_i(M,I)\; p_{j-1}(M,M) + \\ & F_{IM}(i,j-1)\; q_i(I,I)\; p_{j-1}(M,M) \end{aligned} \tag{12}$$

$$\begin{aligned} F_{DG}(i,j) = {} & F_{MM}(i-1,j)\; q_{i-1}(M,D) + \\ & F_{DG}(i-1,j)\; q_{i-1}(D,D) \end{aligned} \tag{13}$$

$$\begin{aligned} F_{MI}(i,j) = {} & F_{MM}(i-1,j)\; q_{i-1}(M,M)\; p_j(M,I) + \\ & F_{MI}(i-1,j)\; q_{i-1}(M,M)\; p_j(I,I) \end{aligned} \tag{14}$$

In an analogous manner our Backward algorithm recursively computes each $B_{MM}(i,j)$ starting from the bottom right of the matrix.

$$\begin{aligned} B_{MM}(i,j) = {} & \\ & 1 \\ + \quad & B_{MM}(i+1,j+1)\; \rho_{\mathrm{aa}}(q_{i+1},p_{j+1})\; q_i(M,M)\; p_j(M,M) \\ + \quad & B_{GD}(i,j+1)\; p_j(M,D) \\ + \quad & B_{IM}(i,j+1)\; q_i(M,I)\; p_j(M,M) \\ + \quad & B_{DG}(i+1,j)\; q_i(M,D) \\ + \quad & B_{MI}(i+1,j)\; q_i(M,M)\; p_j(M,I) \end{aligned} \tag{15}$$

$$\begin{aligned} B_{GD}(i,j) = {} & \\ & B_{MM}(i+1,j+1)\; \rho_{\mathrm{aa}}(q_{i+1},p_{j+1})\; q_i(M,M)\; p_j(D,M) + \\ & B_{GD}(i,j+1)\; p_j(D,D) \end{aligned} \tag{16}$$

$$\begin{aligned} B_{IM}(i,j) = {} & \\ & B_{MM}(i+1,j+1)\; \rho_{\mathrm{aa}}(q_{i+1},p_{j+1})\; q_i(I,M)\; p_j(M,M) + \\ & B_{IM}(i,j+1)\; q_i(I,I)\; p_j(M,M) \end{aligned} \tag{17}$$

$$\begin{aligned} B_{DG}(i,j) = {} & \\ & B_{MM}(i+1,j+1)\; \rho_{\mathrm{aa}}(q_{i+1},p_{j+1})\; q_i(D,M)\; p_j(M,M) + \\ & B_{DG}(i+1,j)\; Q_i(D,D) \end{aligned} \tag{18}$$

$$\begin{aligned} B_{MI}(i,j) = {} & \\ & B_{MM}(i+1,j+1)\; \rho_{\mathrm{aa}}(q_{i+1},p_{j+1})\; q_i(M,M)\; p_j(I,M) + \\ & B_{MI}(i+1,j)\; q_i(M,M)\; p_j(I,I) \end{aligned} \tag{19}$$

Once Forward and Backward functions are calculated, the posterior probabilities can be calculated according to eq. (8).

2 CALCULATION OF EFFECTIVE NUMBER OF SEQUENCES

The effective number of seqences at column i of a multiple alignment is calculated on the subalignment A_i formed by all sequences with a resdiue in column i and by all columns with at most 10% terminal gaps in these sequences. A terminal gap is a gap that lies either to the left or to the right of the entire sequence. For each column j of M_i we calculate amino acid frequencies $p_i(a)$, using the Henniоff sequence weighing scheme. Then the number of effective sequences is

$$N_{\mathrm{eff}}(i) = \exp\left(-\frac{1}{L_i}\sum_{j in M_i}\sum_{a=1}^{20} p_i(a)\log\; p_i(a)\right) . \tag{20}$$

Here, L_i is the number of columns in M_i.

3 SUPPLEMENTARY FIGURES

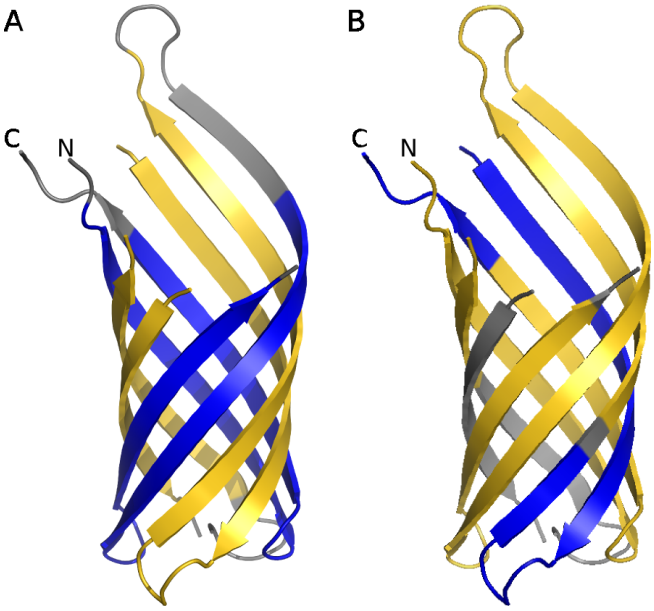

Fig. 1. Structures of outer membrane protein A (OmpA). **(A)** HHrepID correctly identifies three $\beta\beta$-hairpin repeats and two velcro strands with a P-value of 10^{-7}. **(B)** RADAR predicts only three repeats whose borders are spuriously placed inside β-strands. The repeat units in OmpA are so highly diverged, that neither HMMER nor TRUST are able to detect their repeat signature.

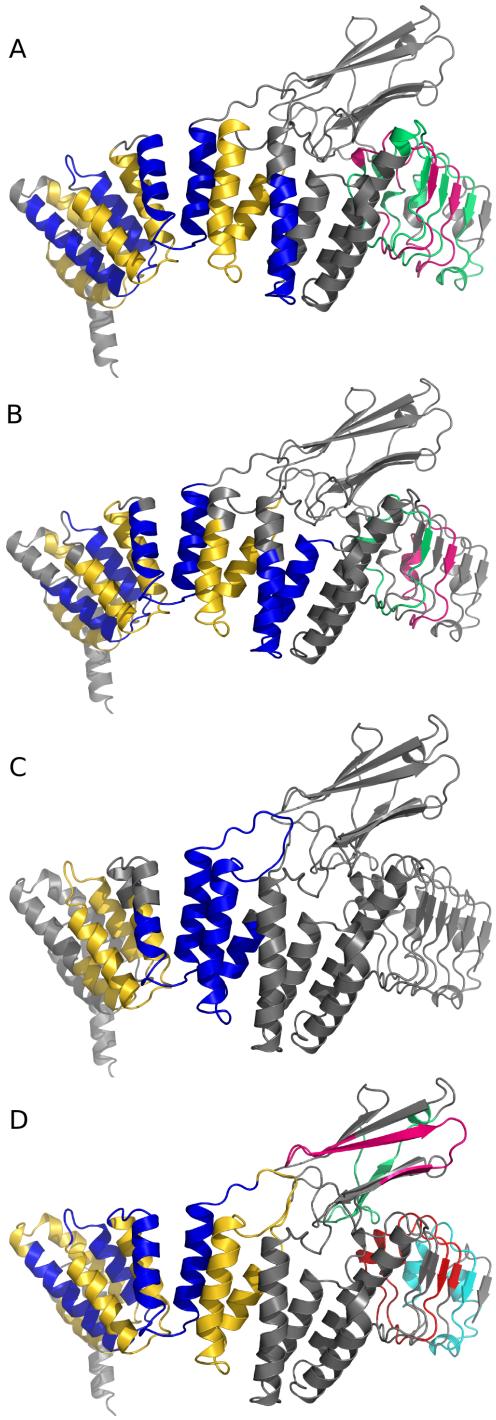

Fig. 2. Structures of RAB geranylgeranyltransferase alpha subunit (PDB identifier: 1DCE) consisting of six prenyltransferase α subunit repeats and five Leucine-rich repeats. **(A)** HHrepID correctly identifies all six prenyltransferase repeats including the inserted domain after the fifth α-hairpin repeat and all five Leucine-rich repeats at the C-terminus. **(B)** HMMER detects both repeat families but still misses three of the five Leucine-rich repeats. **(C)** TRUST identifies two repeats in the prenyltransferase repeat region but their repeat lengths and borders are clearly wrong. **(D)** RADAR is able to identify four prenyltransferase repeats but predicts four wrong repeat units in the inserted domain and in the C-terminal region

REFERENCES

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998) *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge.

Kschischo, M. and Lssig, M. (2000) Finite-temperature sequence alignment. *Pac Symp Biocomput*, pp. 624–635.

Miyazawa, S. (1995) A reliable sequence alignment method based on probabilities of residue correspondences. *Protein Eng*, **8**, 999–1009.

Mückstein, U., Hofacker, I. L. and Stadler, P. F. (2002) Stochastic pairwise alignments. *Bioinformatics*, **18 Suppl 2**, 153–160.

Söding, J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.