BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 6 2008, pages 807–814
doi:10.1093/bioinformatics/btn039

Sequence analysis

De novo identification of highly diverged protein repeats by probabilistic consistency

A. Biegert[1,2] and J. Söding[1,2,*]

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

Received on November 17, 2007; revised on January 2, 2008; accepted on January 24, 2008

Advance Access publication February 1, 2008

Associate Editor: Limsoon Wong

ABSTRACT

Motivation: An estimated 25% of all eukaryotic proteins contain repeats, which underlines the importance of duplication for evolving new protein functions. Internal repeats often correspond to structural or functional units in proteins. Methods capable of identifying diverged repeated segments or domains at the sequence level can therefore assist in predicting domain structures, inferring hypotheses about function and mechanism, and investigating the evolution of proteins from smaller fragments.

Results: We present HHrepID, a method for the *de novo* identification of repeats in protein sequences. It is able to detect the sequence signature of structural repeats in many proteins that have not yet been known to possess internal sequence symmetry, such as outer membrane β-barrels. HHrepID uses HMM–HMM comparison to exploit evolutionary information in the form of multiple sequence alignments of homologs. In contrast to a previous method, the new method (1) generates a multiple alignment of repeats; (2) utilizes the transitive nature of homology through a novel merging procedure with fully probabilistic treatment of alignments; (3) improves alignment quality through an algorithm that maximizes the expected accuracy; (4) is able to identify different kinds of repeats within complex architectures by a probabilistic domain boundary detection method and (5) improves sensitivity through a new approach to assess statistical significance.

Availability: Server: http://toolkit.tuebingen.mpg.de/hhrepid; Executables: ftp://ftp.tuebingen.mpg.de/pub/protevo/HHrepID

Contact: soeding@lmb.uni-muenchen.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

About 25% of all eukaryotic protein sequences contain repeating amino acid segments (Marcotte *et al.*, 1999). The percentage of repeat-containing proteins grows with the complexity of the organism, with repeat proteins being particularly abundant in multicellular organisms (Bjorklund *et al.*, 2006). In vertebrates, for example, tandemly arranged repeats often serve as a structural framework for the formation of protein–protein interactions (Kobe and Kajava, 2001; Li *et al.*, 2006). Furthermore, assemblies of repeats are readily evolvable and provide an organism with opportunities to easily expand its repertoire of cellular functions (Street *et al.*, 2006).

We would like to predict repeats from protein sequences for the following reasons: (1) It can help to elucidate the domain structure of multi-domain proteins by determining the boundaries of domains with internal repeats or by detecting the presence of duplicated structural domains. This facilitates the application of subsequent sequence analysis methods and can help to design constructs for X-ray crystallography. (2) For proteins without any known homologs, the identification of repeats may give hints to their fold or family. (3) Since duplication is an important mechanism to generate new folds, the determination of protein repeats may yield insights into the origin of protein folds (Lupas *et al.*, 2001; Söding and Lupas, 2003).

There are three classes of methods to detect repeats in protein sequences. The first is specialized in detecting repeats in fibrous proteins and does not allow for insertions within (Coward and Drablos, 1998) or between repeat units (Gruber *et al.*, 2005; Lupas *et al.*, 1991; McLachlan and Stewart, 1976; Newman and Cooper, 2007).

The second class utilizes a database of single repeat units in the form of sequence profiles or profile hidden Markov models (HMMs) that have been compiled from known repeat families. These profiles are compared one by one to the query sequence. To be able to detect multiple instances of a particular repeat type, more than one hit to a repeat profile is allowed. The well-known HMMER/Pfam package (Eddy, 1998; Sonnhammer *et al.*, 1998) as well as REP (Andrade *et al.*, 2000) and Mocca (Notredame, 2001) belong to this group.

The third class of methods does not rely on a priori knowledge about repeat families. Instead, these methods detect internal sequence symmetries by comparing the protein sequence to itself. Six methods belong to this class: internal repeat finder (Sonnhammer *et al.*, 1998), PROSPERO (Mott, 2000), REPRO (Heringa and Argos, 1993), RADAR (Heger and Holm, 2000), TRUST [the successor of REPRO, Szklarczyk and Heringa (2004)], and the HHrep server (Söding *et al.*, 2006). With the exception of HHrep, all methods utilize sequence–sequence comparison to find suboptimal

*To whom correspondence should be addressed.

self-alignments. The HHrep server is a straightforward implementation of HMM–HMM comparison that exploits evolutionary information in the form of homologs to the query. RADAR and TRUST are the only tools that build a repeat profile to determine exact repeat borders and thereby extract a multiple alignment of repeats.

Only HHrep and TRUST explicitly make use of consistency (also termed *transitivity* in this context), a concept that has led to significant improvements in multiple sequence alignment (Notredame *et al.*, 2000). Owing to consistency, TRUST and HHrep can find additional suboptimal self-alignments that were either missed or previously deemed insignificant. Consequently, TRUST and HHrep are the methods that have been reported to be most sensitive to date (Söding *et al.*, 2006).

Here, we present an HMM-based *de novo* method with several novel algorithmic improvements. First, we extend the *maximum expected accuracy* (MAC) algorithm (Holmes and Durbin, 1998), which maximizes the sum of *posterior probabilities* in the alignment, to the case of local HMM–HMM alignment. Second, we pursue a fully probabilistic approach to consistency through a novel merging procedure based on posterior probabilities. Third, we automatically detect domain boundaries allowing for the identification of different repeat types within complex multi-domain architectures. Due to its extreme sensitivity, the presented method is able to detect for the first time with high significance very divergent repeat patterns in many outer membrane β-barrels (OMPs), which points to their origin by amplification of a single β-hairpin (Remmert, Biegert *et al.*, to be published).

2 MATERIALS AND METHODS

A flow diagram with all steps of the HHrepID repeat detection algorithm is given in Figure 1. The individual steps will be explained in the following sections.

2.1 Generation of query HMM

In order to construct an HMM of the query sequence, the HHrepID algorithm requires an alignment of protein sequences as input. HHrepID utilizes the `buildali.pl` PERL script of the HMM–HMM comparison suite HHsearch (Söding, 2005) to construct an alignment of homologs with PSI-BLAST [Altschul *et al.* (1997), `buildali.pl` was run with default parameters and is available at ftp://ftp.tuebingen.mpg.de/pub/protevo/HHsearch]. Before the query HMM is generated by HHrepID, the input alignment is filtered to a maximum bit per column score of 0.3 between the query sequence and its homologs to reduce the influence of non-homologous fragments and wrongly aligned homologs.

2.2 Posterior probabilities

To detect sequence signatures of repeats in a query protein, HHrepID uses local HMM–HMM self-comparison. In addition to calculating the alignment with the maximum score with the Viterbi algorithm, HHrepID computes posterior probabilities with the Forward/Backward algorithm. Given two sequences x and y, the posterior probability $P(x_i \diamond y_j \mid x,y)$ quantifies the probability that residue i in sequence x is aligned to residue j in sequence y (Miyazawa, 1995). This approach was extended to the case of *local* sequence–sequence alignment by Mückstein *et al.* (2002). In order to be able to detect local repeat patterns, we generalize the concept of posterior probabilities to the case of *local HMM–HMM* comparison. Furthermore, we introduce a *random model* in the calculation of posterior probabilities, a measure that was found to lead to significant gains in sensitivity for Viterbi alignment [HMM-to-sequence comparison: Durbin *et al.* (1998), HMM-to-HMM comparison: Söding, unpublished]. The details about the calculation of posterior probabilities for local HMM–HMM comparison can be found in the Supplementary Material.

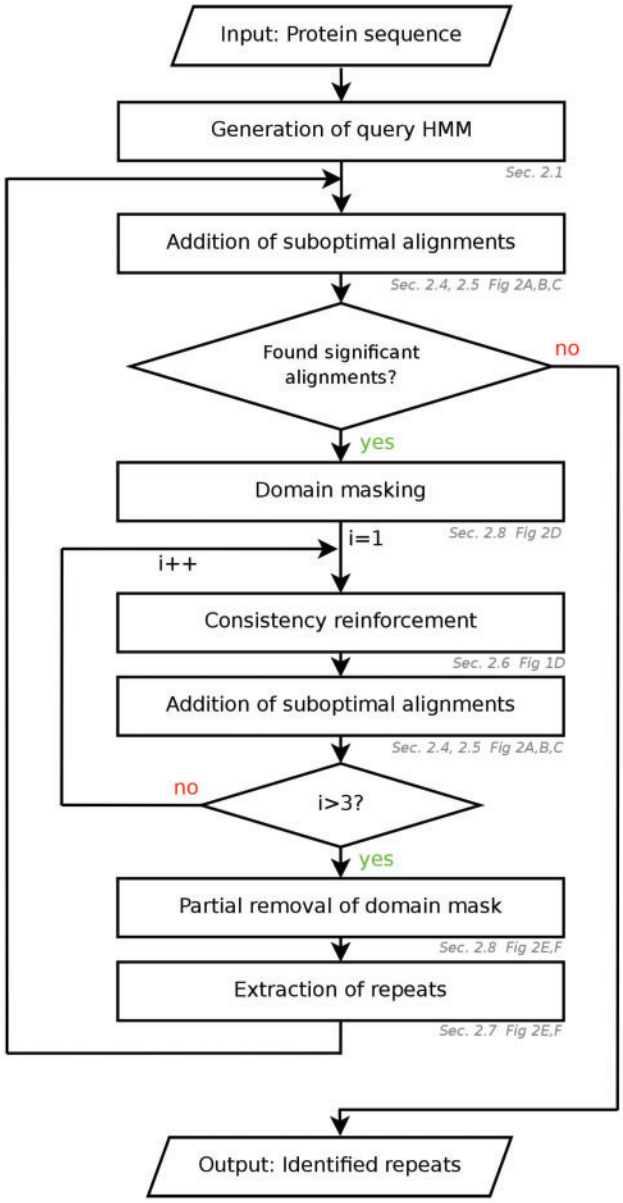

Fig. 1. Flowchart illustrating the main steps of the HHrepID repeat detection algorithm. References at the lower right of each process box indicate the section describing the step in detail and give the respective subfigures in Figure 2.

2.3 Maximum accuracy alignment

To derive an alignment from a posterior probability matrix, Holmes and Durbin (1998) proposed a maximum accuracy (MAC) algorithm.

A MAC alignment π maximizes the expected number of correctly aligned pairs of residues:

$$\mathcal{A}(\pi) = \sum_{(i,j)\in\pi} P(M_i^q \diamond M_j^p) \rightarrow \max \ . \quad (1)$$

Here, $P(M_i^q \diamond M_j^p)$ denotes the posterior probability of match state i in HMM q to be aligned to match state j in HMM p [see Söding (2005) for details about HMM–HMM alignment]. Because $P(M_i^q \diamond M_j^p)$ is always positive, MAC alignments maximizing $\mathcal{A}(\pi)$ are always global. Since we need *local* alignments, we introduce an additional probability threshold parameter T that governs the greediness of the MAC alignment. Our local MAC alignment should maximize the expected sum of posterior probabilities minus probability T per aligned pair along the alignment path. This new type of alignment score will generally give a more accurate alignment than the Viterbi algorithm and has the additional advantage of a tunable 'greediness' parameter. The method to derive our local maximum accuracy alignment is a modified version of the method proposed by Holmes and Durbin (1998), which uses the posterior probabilities as substitution scores:

$$A(i,j) = \max \begin{cases} 0, \\ A(i-1,\, j-1) + P(M_i^q \diamond M_j^p) - T, \\ A(i-1,\, j) - T/2, \\ A(i,\, j-1) - T/2. \end{cases} \quad (2)$$

After matrix A has been filled, a standard traceback procedure will produce the best alignment. The cost of $T/2$ associated with the placement of gaps in the alignment ensures that the local MAC alignments are compact. A gap in HMM p cannot be followed directly by a gap in HMM q since the gap penalties would outweigh the mismatch score incurred when the two unaligned regions are shifted onto each other.

Although a MAC alignment path often overlaps with the optimal Viterbi path, there are cases where MAC and Viterbi alignments differ completely. Unlike Viterbi scores, MAC scores are not guaranteed to be distributed according to an extreme value distribution (EVD). This makes the Viterbi algorithm more suited for the calculation of P-values.

2.4 Addition of suboptimal alignments

When compared to itself, a typical repeat protein with n consecutive repeat units will give rise to $2n-1$ suboptimal alignments: one trivial self-alignment and $n-1$ pairs of equivalent alignments. The correct and hence consistent set of suboptimal alignments contains the complete information about length and spacing of repeats. To find a set of suboptimal alignments that is maximally accurate and consistent our method uses posterior probabilities calculated with the Forward/Backward algorithm.

Each search for suboptimal alignments starts with the calculation of a Viterbi alignment. If the P-value of the Viterbi alignment is above a specified threshold the current search for suboptimal alignments is terminated. Otherwise, HHrepID proceeds to calculate a posterior probability matrix P with the Forward/Backward algorithm. The posterior probabilities are also recorded in a total posterior probability matrix P^{tot}, which has been initially set to the identity matrix. After calculating the posterior probabilities P_{ij} for the next suboptimal alignment, we update the *total* probability matrix by taking the maximum (Fig. 2C):

$$(P_{ij}^{\text{tot}})' = \max(P_{ij}, P_{ij}^{\text{tot}}). \quad (3)$$

Matrix P will normally contain only the trace of the next best suboptimal alignment instead of all remaining suboptimal alignments. By using max in Equation (3) we ensure that the total posterior probability matrix P^{tot} gathers the probabilities of all suboptimal alignment traces that have been found so far. The newly calculated posterior probabilities P_{ij} not only contribute to the total posterior matrix but are also used as substitution scores in the calculation of a MAC alignment. The cells covered by the MAC alignment are masked to be able to find the next alignment, similar to the algorithm by Waterman and Eggert (1987). Figure 2A shows the first suboptimal MAC alignment in the repeat protein 1DCE and the posterior probability matrix P_{ij} from which it was computed. The Viterbi, Forward/Backward and MAC computation steps are repeated until no further significant suboptimal alignments are found (Fig. 2B).

2.5 Suppression of spurious alignments of repeats

Spurious suboptimal alignments are caused by chance similarities between non-homologous regions within repeat units. The insignificant score of one such match may then get multiplied by the number of repeat units minus one. This problem becomes more severe as the number of repeats increases. It is therefore difficult to discriminate between true and spurious self-alignments of a repeat protein on the basis of their P-values. Spurious alignments are normally *inconsistent* with the already detected correct suboptimal alignments. The suppression of spurious alignments works by restricting the search for the next suboptimal alignment to those cells in the dynamic programming matrix that are approximately *consistent* with already found suboptimal alignments. Suppose we have detected a suboptimal alignment in which residue i is aligned to j and j is aligned to k. Then, we can infer that there has to be a second suboptimal alignment which aligns residue i to k. Such an alignment would be *consistent* with the already found suboptimal alignment. This concept of consistency can also be formulated in terms of posterior probabilities. Do *et al.* (2005) employ a *probabilistic consistency transformation* in the multiple alignment program ProbCons. In HHrepID, we utilize a similar approach and mask cells (i,k) that are inconsistent with already detected suboptimal alignments:

$$\{(i,k) : \frac{1}{r-1} \sum_{j=1}^{L} P_{ij}^{\text{tot}} P_{jk}^{\text{tot}} < 0.001\} \ . \quad (4)$$

Here, $1/(r-1)$ is simply a scaling factor to ensure that the results of the consistency transformation stay below 1 when searching for the r-th suboptimal alignment. The low value of the threshold ensures that only cells that are completely inconsistent with the posterior probabilities of already found suboptimal alignments are masked. (We mask cells by forcing their score to $-\infty$ and their probability to zero.) As an example, brown regions in Figure 2B represent cells that are inconsistent with the posterior probabilities computed during the search for the first suboptimal alignment (Fig. 2A). These cells have been excluded from the search for the second suboptimal alignment.

In addition to this masking procedure, we can utilize the repeat length w to further suppress spurious suboptimal alignments. The repeat length w is estimated by the median distance of the best scoring MAC alignment to the main diagonal (see Fig. 2A). Then, in addition to Equation (4), we mask all cells within $w/2$ positions from the already found MAC alignments.

2.6 Consistency reinforcement

After no further significant Viterbi alignment can be found (Fig. 2C) we improve the consistency of the total posterior matrix. This may also permit to reconstruct traces of missing suboptimal alignments that are consistent with the already detected suboptimal alignments. One way to amplify the consistency would be the repeated application of Do's probabilistic consistency transformation $p'_{ik} = \text{const} \times \sum_j p_{ij} p_{jk}$. However, this will cause the posterior matrix either to converge to the null matrix (if the largest eigenvalue λ of the initial posterior

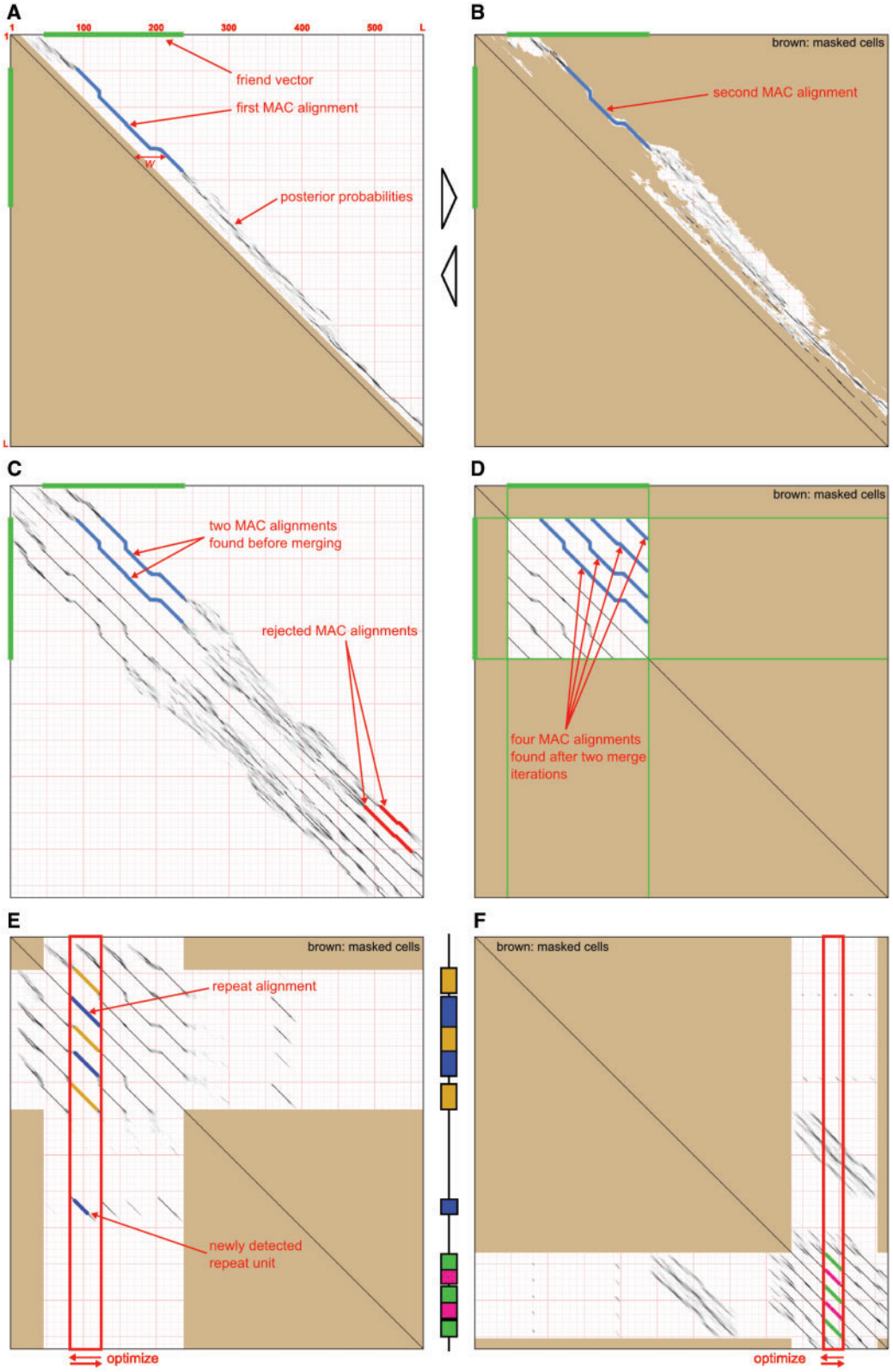

Fig. 2. Basic steps of the HHrepID algorithm exemplified for the protein 1DCE consisting of six α-hairpin repeats and five Leucine-rich repeats. See main text for details. (**A**) Generate next suboptimal alignment. (**B**) Mask inconsistent regions. (**C**) Accumulate posteriors into total posterior matrix. (**D**) Mask, merge HMM and iterate steps A–C. (**E**) Determine alignment of repeats of type A. (**F**) Determine alignment of repeats of type B.

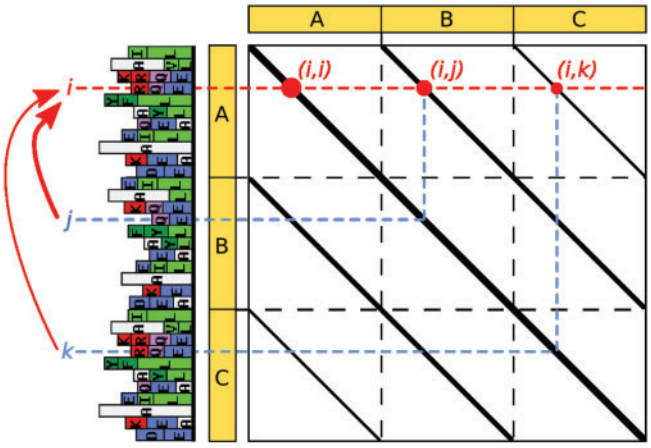

Fig. 3. Consistency reinforcement by merging the query HMM with itself. The figure shows a repeat protein with three repeats A, B, C, its amino acid profile on the left for illustration, and the posterior probability matrix containing the trivial self alignment and two suboptimal alignments in the upper triangular matrix. The new amino acid emission probabilities at positions i are a weighted mixture of the amino acid emission probabilities at position i,j and k. The size of the circular marks indicates the weights.

matrix is <1) or to explode (if $\lambda > 1$). Hence Do's transformation does not lead to convergence.

To reinforce consistency of the posterior matrix, HHrepID utilizes a merging procedure that mixes the amino acid emission probabilities of homologous columns in the query HMM with each other. If position i has been found to be aligned with positions j and k in suboptimal alignments we want the emission probabilities at position i to be a weighted mixture of the emission probabilities at i, j and k. Each column j should contribute to the mixture according to the certainty that position i and j are homologous, which is simply the posterior probability P_{ij}^{tot}:

$$p_i'(a) = \text{const} \times \sum_{j=1}^{L} P_{ij}^{\text{tot}} N_{eff}^{M}(j) p_j(a) \tag{5}$$

where $N_{eff}^{M}(j)$ is the number of effective sequences going through state M_j. [We refer to the supplementary material for the calculation of $N_{eff}^{M}(j)$.] Figure 3 illustrates the merging step for an idealized repeat protein with three repeats and two suboptimal alignments. Note that $P_{ij}^{\text{tot}} = 1$ for $i = j$. We can merge the transition probabilities $p_j(X \to Y)$ of the HMM with itself in an analogous manner:

$$p_i'(X \to Y) = \text{const} \times \sum_{j=1}^{L} P_{ij}^{\text{tot}} N_{eff}^{X}(j) p_j(X \to Y) \tag{6}$$

where $X \to Y$ stands for the transitions $M \to M$, $M \to I$, $M \to D$, $I \to M$ etc. By merging the HMM with itself, HHrepID probabilistically incorporates consistency information in the updated HMM. We observed that in most cases when this new HMM is again compared to itself, a much cleaner repeat pattern emerged and regions of washed out probability density converged to the most probable alternative. Often, this allowed us to identify previously undetected self-alignments. As an example, Figure 2D illustrates how the repeat pattern of 1DCE improves after two merge iterations. The presented merge routine always converges, generally after three merge iterations, the default setting in HHrepID. The improved repeat pattern recorded in the total posterior matrix serves as the main input for the last step in the HHrepID algorithm: the extraction of repeat units.

2.7 Extraction of repeats

Let us now explain the extraction of repeat units for the simpler case where there is just *one* family of repeats. After the last search for suboptimal alignments has been performed, HHrepID determines the repeat borders of individual repeats. To do so, the algorithm requires two kinds of information: (1) The traces of all suboptimal alignments in the form of the converged total posterior probability matrix P^{tot}, and (2) the last estimate of repeat length w (see Section 2.5).

First, HHrepID calculates the position of the representative repeat that is most similar to all other repeats with the help of a sliding window of w consecutive columns in the posterior matrix. The window should be positioned over the best conserved columns in the matrix. To quantify how well a particular position j is conserved throughout repeats, we define

$$c(j) = \sum_{i=1}^{L} (P_{ij}^{\text{tot}})^2.$$

This score $c(j)$ is high for well-focused posteriors and spread out for columns with a fluffy distribution of posteriors because the posterior probabilities are summed with their *squared* value. Furthermore, we would like to choose the repeat borders of the representative repeat such that the middle part consists of well-focused posteriors whereas the linkers may show a more smeared out probability distribution. This reflects our assumption that repeats are generally well conserved in the core and linked by less-conserved regions. Every column score $c(j)$ is therefore weighted according to its position in the sliding window with weights in the center being higher than at its ends. More precisely, the weight $w(d)$ for position d within the window is given by a simple triangular function

$$w(d) = \begin{cases} d - \frac{1}{2}, & d \le \frac{w}{2} \\ w - d + \frac{1}{2}, & d > \frac{w}{2}. \end{cases} \tag{7}$$

Among all possible windows, the one with the highest sum of weighted column scores is chosen as representative repeat:

$$S(k) = \sum_{d=0}^{w-1} w(d)\, c(k+d) \to \max. \tag{8}$$

The dot plot in Figure 2E illustrates the positioning of the representative repeat with its borders placed over smeared out regions.

Once the position of the representative repeat is fixed, the MAC alignments with all other repeats are calculated with the method of Equation (2), using the values from the existing total posterior matrix. After all repeat instances have been identified in this way, HHrepID utilizes the pairwise alignments between each repeat unit and the representative repeat to infer a multiple alignment of all repeat units.

2.8 Repeats in multi-domain proteins

We would like to be able to extract different kinds of repeats in complex architectures. However, the algorithm as it has been presented so far is only suited for single-domain proteins containing a single type of repeats. Through the consistency reinforcement iterations, chance similarities may amplify themselves and cause suboptimal alignments to grow into non-homologous regions. Therefore, we would like to mask all regions in the HMM that presumably do not contain repeats or contain repeats that belong to a different family than those that have given rise to the first suboptimal alignment. The algorithm will then process the unmasked regions as described in the previous sections and extract all repeats of one particular type (Fig. 2E). After all repeats have been identified, the corresponding regions in the query HMM are *permanently* masked before the next repeat alignment is calculated (Fig. 2F). This procedure is then repeated until no further significant repeats can be identified.

To determine the regions that have to be masked, we define a Boolean vector (*friend vector*). It keeps track of all positions in the query that belong to the same type of repeats as those matched in the first suboptimal MAC alignment (see green positions in Figure 2A). In this way, the first MAC alignment determines the type of repeats to be extracted in the following search rounds. If a newly detected MAC alignment has a certain minimum overlap (five residues) with the marked vector positions, these positions are added to the friend vector and Equation (3) is applied. Otherwise the suboptimal alignment and its posterior probability matrix is rejected. However, previously rejected alignments can later be accepted if they overlap with a later friend vector. Figure 2C shows two suboptimal alignments (red), which belong to Leucine-rich repeats in the protein 1DCE, that have been rejected because they did not overlap with the friend vector (green). After the initial search for suboptimal alignments has been completed, all unmarked positions in the friend vector are masked in the dynamic programming matrix and thereby excluded from the following search rounds and consistency reinforcement iterations (see brown regions in Fig. 2D).

Since the masking procedure has to be fairly conservative, regions that have been masked might contain repeats which are so highly diverged that they could not be picked up by suboptimal alignments in the initial search round. By *partially* removing the domain mask just before the last search for suboptimal alignments (Figure 2E and F), the detection of so far undetected repeat units is often possible through the improved consistency (Fig. 2E).

2.9 Calculation of *P*-values

We calculate two kinds of *P*-values to assess the statistical significance of HHrepID's repeat detection results: Repeat group *P*-values, which refer to the whole group of homologous repeat units found together, and *P*-values for individual repeat units. To compute the repeat group *P*-value, we search a calibration database of unrelated proteins and fit the parameters of the extreme value distribution (EVD) to the scores. The repeat group *P*-value is simply the *P*-value of the best suboptimal Viterbi self-alignment given this EVD. The repeat-specific *P*-value is determined during the repeat extraction stage: We first build a repeat profile HMM of repeat length w by merging [Equations (5) and (6)] all regions within the repeat extraction window (red box in Figure 2E) that do not belong to the repeat to be evaluated. In this way, we obtain a profile HMM that represents all but one consistently mixed repeat units. With this repeat profile we search our calibration database and fit parameters of the EVD, allowing us to calculate the repeat *P*-value of the alignment between the original, unmerged repeat unit and the repeat profile.

3 RESULTS AND DISCUSSION

We evaluate the repeat detection performance of HHrepID and compare it to the newest *de novo* methods, TRUST and RADAR, as well as to the database-dependent method HMMER/Pfam (version 2.3.2). The HHrep server is not included because it does not calculate explicit repeat alignments and borders. The benchmark dataset consists of all protein chains in the protein databank PDB [Berman *et al.* (2000), revision February 2007], filtered to a maximum sequence identity of 20% (6992 chains). We used default settings for all tested methods and the global Pfam library. (The fragment library is not suitable to detect repeats since local alignment would not enforce overlap of detected domain or repeat fragnments.) For HHrepID, parameter T [Equation (2)] is set to 0.5, the total repeat *P*-value threshold for repeat families to 10^3, and the *P*-value threshold for suboptimal alignments to 0.1.

For each method, we would like to determine the number of false positive and true positive repeat proteins detected. Ideally, we would compare the repeat alignments with a gold standard method, e.g. one based on structural alignment of repeat units. Unfortunately, although one such method has been developed (Murray *et al.*, 2004), a tool is not available. We therefore use the reported pairwise alignments between repeat units instead to decide which proteins will be considered as true positive repeat proteins. Predicted repeat alignments that yield a good structural alignment are defined as correct. To classify a protein as true positive repeat protein, we require that at least 50% of predicted repeats in the protein have a correct alignment to at least one other repeat unit. We define an alignment as correct if its TMscore (Zhang and Skolnick, 2004) is greater or equal to 0.4 ($\text{TMscore} \in [0,1]$), which is the threshold for significant structural similarity given by Zhang. (We checked that results are very similar for a threshold of 0.5.) We ignore all repeats that are shorter than 15 residues because TMscore is unsuited for such short repeats. By conditioning the acceptance of a repeat alignment on the alignment quality our benchmark not only assesses performance in detecting repeat proteins but also measures the quality of the underlying repeat alignments.

Figure 4A plots the number of true positive repeat-containing proteins against the number of false positive proteins detected above a threshold significance. As significance measure we use the *P*-value of the best suboptimal self-alignment (HHrepID), the reported *E*-value (HMMER), the repeat family *P*-value (TRUST) and the repeat family score (RADAR). At a cumulative error rate of 10% HHrepID detects about thirty times more repeat proteins than RADAR, about twice as many as TRUST, and about 20% more than HMMER. Furthermore, HHrepID and TRUST show a very low error rate at high significance levels: They are able to detect 120 (HHrepID) and 80 (TRUST) true positives before identifying the first false positive. In contrast, HMMER reports several wrong alignments with very significant *E*-values. Although HMMER correctly picks up the sequence signature of duplicated domains in these cases, the quality of the constructed repeat alignments is below the cut-off threshold of 0.4.

This benchmark measures how well a method can differentiate between repetitive and non-repetitive proteins. It would also be helpful to evaluate true positive versus false positive rates on the level of individual repeat units since in practical applications the exact repeat positions and borders are needed. We therefore perform a second, repeat-specific test in which a predicted repeat unit is judged as true positive if it takes part in at least one alignment with TMscore greater than 0.4. As score we use repeat probabilities (HHrepID Prob, see below), repeat *P*-values (HHrepID *P*-value), and repeat *E*-values (HMMER). Since TRUST and RADAR do not report repeat-specific significance scores we rank results by the repeat family *P*-value (TRUST) and repeat family score (RADAR).

Figure 4B shows the results of the repeat unit level benchmark. The results are qualitatively similar to Figure 4A: HHrepID is able to identify about twenty times more repeats than RADAR, about twice as many repeats as TRUST and

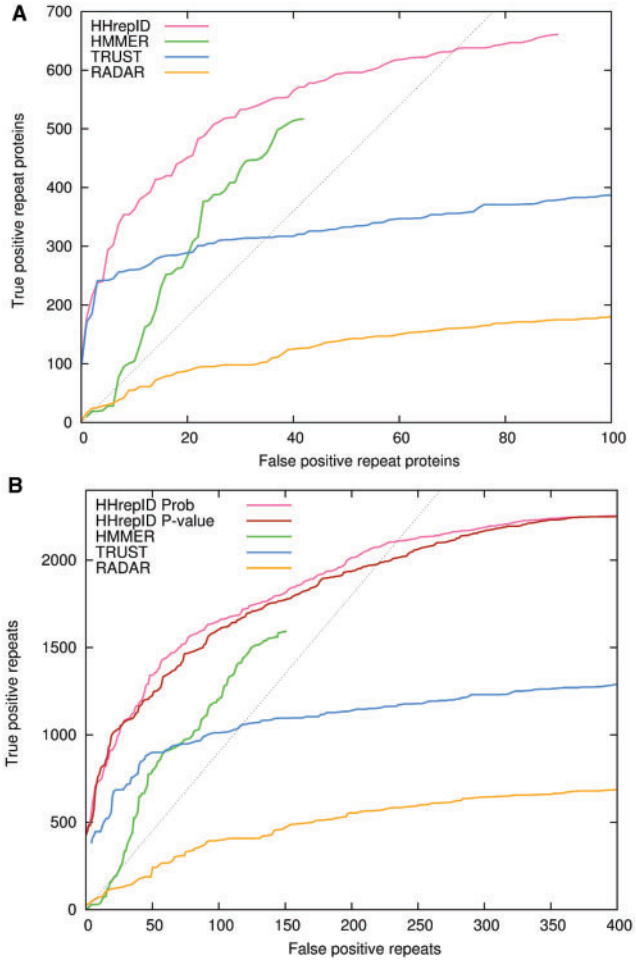

Fig. 4. **(A)** Number of correctly detected repeat proteins (true positives) versus number of false positives above a given significance score. The dotted line indicates an error rate of 10%. **(B)** Number of correctly detected repeat units (true positives) versus number of false positives above a given significance score.

about 30% more than HMMER at a cummulative error rate of 10%. The calculation of repeat-specific probabilities only slightly enhances the sensitivity (HHrepID Prob). Probabilities are calculated by kernel density estimation in a three-dimensional space formed by the negative log of repeat *P*-value, repeat length, and the length-normalized MAC score obtained from Equation (2). (We use a Gaussian kernel and 2-fold crossvalidation on our benchmark dataset.)

When interpreting the results in Figure 4A and B, one has to keep in mind that, in contrast to the next best method HMMER, HHrepID does not rely on a priori knowledge in the form of a precompiled database of repeat families. On less well studied sequences HMMER is likely to perform worse than in our benchmark, whereas HHrepID's performance will stay the same. It should also be noted that we did not remove sequences with trivial repeat signatures. The difference between HHrepID and the other tools is therefore expected to be much more pronounced when repeats have significantly diverged in sequence.

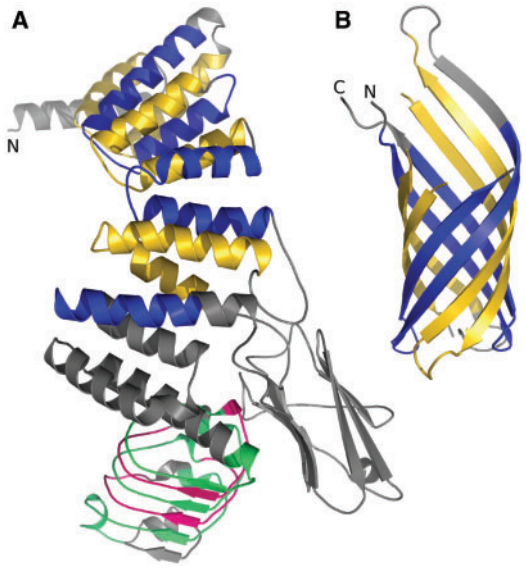

Fig. 5. **(A)** Structure of RAB geranylgeranyltransferase alpha subunit (PDB identifier: 1DCE). Repeat units identified by HHrepID are highlighted according to colors used in Figure 2e and f. **(B)** Structure of outer membrane protein A. HHrepID correctly identifies three $\beta\beta$-hairpin repeats and two velcro strands with a *P*-value of 10^{-7}.

To demonstrate HHrepID's ability to detect repeats in complex architectures we analyze RAB geranylgeranyltransferase (1DCE, Fig. 5A) which also served as example in Figure 2. It consists of six prenyltransferase α subunit repeats and five Leucine-rich repeats. 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. For comparison, HMMER detects both repeat families but still misses three of the five Leucine-rich repeats (Figure S2B). RADAR is able to identify four prenyltransferase repeats but predicts four wrong repeat units in the inserted domain and in the C-terminal region (Figure S2D). TRUST identifies two repeats in the prenyltransferase repeat region but their repeat lengths and borders are clearly wrong (Figure S2C).

We further applied HHrepID to outer membrane β barrels which are formed of between 4 and 11 $\beta\beta$-hairpins in a barrel-shaped arrangement. The structure with the hairpin repeats looks fairly regular, but until now a repeat signature in sequence had not been identified (see Fig. S1B for results of RADAR) (Neuwald *et al.*, 1995). HHrepID is able to detect a clear and unambiguous $\beta\beta$-hairpin repeat in half of all outer membrane β barrels in the PDB. As an example, Figure 5B shows results for OmpA where HHrepID even correctly identifies the two velcro strands at the N and C-terminus.

4 CONCLUSION

During the development of HHrepID we devised several algorithmic improvements for the detection of diverged repeats, the placement of repeat borders, and the analysis of complex domain architectures. Several of these algorithmic advances are not restricted to the use in repeat detection: we generalized the Forward–Backward algorithm to the case of local

HMM–HMM alignment with a random model and use it to calculate posterior probabilities, allowing the quantification of the local reliability of an alignment. Our modified maximum accuracy (MAC) alignment algorithm contains a tunable greediness parameters with which we can continuously switch between local and global alignment. This local MAC algorithm has been implemented in the HMM–HMM comparison suite HHsearch (Söding, 2005), version 1.5.0. Our probabilistic consistency reinforcement procedure that acts on profiles rather than posterior probabilities could also be transferred to multiple sequence alignment, where consistency has been used to suppress early alignment errors during progressive alignment.

Due to HHrepID's increased sensitivity we are able to detect the sequence signature of repeats in several ancient folds that have not yet been known to possess internal sequence symmetry, such as TIM barrels and outer membrane β-barrels. Analysis with HHrepID confirms earlier results that TIM barrels evolved by amplification of quarter barrel fragments (Nagano *et al*., 1999; Söding *et al*., 2006) and not half barrels (Lang *et al*., 2000). These results further support the 'ancient peptide hypothesis', which posits that modern proteins arose by recombination and fusion from a small set of fragments (ancient peptides) (Lupas *et al*., 2001; Söding and Lupas, 2003). We believe that the repeats detected with HHrepID in OMPs and TIM barrels are evolutionary remnants of these ancient fragments.

ACKNOWLEDGEMENTS

We thank Michael Remmert, Christian Mayer and Oliver Kohlbacher for stimulating discussions. We are particularly grateful to Andrei Lupas for initiating this work and giving ample support and advice. Financial support by the Max-Planck-Society and partial support by the Center for Integrated Protein Science Munich (CIPSM) is gratefully acknowledged.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res*., **25**, 3389–3402.

Andrade,M.A. *et al*. (2000) Homology-based method for identification of protein repeats using statistical significance estimates. *J. Mol. Biol*., **298**, 521–537.

Berman,H.M. *et al*. (2000) The Protein Data Bank. *Nucl. Acids Res*., **28**, 235–242.

Bjorklund,A.K. *et al*. (2006) Expansion of protein domain repeats. *PLoS Comput. Biol*., **2**, e114.

Coward,E. and Drablos,F. (1998) Detecting periodic patterns in biological sequences. *Bioinformatics*, **14**, 498–507.

Do,C.B. *et al*. (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res*., **15**, 330–340.

Durbin,R. *et al*. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Gruber,M. *et al*. (2005) REPPER – repeats and their periodicities in fibrous proteins. *Nucl. Acids Res*., **33**, 239–243.

Heger,A. and Holm,L. (2000) Rapid automatic detection and alignment of repeats in protein sequences. *Proteins Struct. Funct. Genetics*, **41**, 224.

Heringa,J. and Argos,P. (1993) A method to recognize distant repeats in protein sequences. *Proteins Struct. Funct. Genetics*, **17**, 391–341.

Holmes,I. and Durbin,R. (1998) Dynamic programming alignment accuracy. *J. Comput. Biol*., **5**, 493–504.

Kobe,B. and Kajava,A.V. (2001) The leucine-rich repeat as a protein recognition motif. *Curr. Opin. Struct. Biol*., **11**, 725–732.

Lang,D. *et al*. (2000) Structural evidence for evolution of the beta/alpha barrel scaffold by gene duplication and fusion. *Science*, **289**, 1546–1550.

Li,J. *et al*. (2006) Ankyrin repeat: a unique motif mediating protein-protein interactions. *Biochemistry*, **45**, 15168–15178.

Lupas,A. *et al*. (1991) Predicting coiled coils from protein sequences. *Science*, **252**, 1162–1164.

Lupas,A.N. *et al*. (2001) On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J. Struct. Biol*., **134**, 191–203.

Marcotte,E.M. *et al*. (1999) A census of protein repeats. *J. Mol. Biol*., **293**, 151–160.

McLachlan,A.D. and Stewart,M. (1976) The 14-fold periodicity in alpha-tropomyosin and the interaction with actin. *J. Mol. Biol*., **103**, 271–298.

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

Mott,R. (2000) Accurate formula for *P*-values of gapped local sequence and profile alignments. *J. Mol. Biol*., **300**, 649–659.

Mückstein,U. *et al*. (2002) Stochastic pairwise alignments. *Bioinformatics*, **18** (Suppl. 2), 153–160.

Murray,K.B. *et al*. (2004) Toward the detection and validation of repeats in protein structure. *Proteins*, **57**, 365–380.

Nagano,N. *et al*. (1999) Barrel structures in proteins: automatic identification and classification including a sequence analysis of TIM barrels. *Protein Sci*., **8**, 2072–2084.

Neuwald,A.F. *et al*. (1995) Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci*., **4**, 1618–1632.

Newman,A.M. and Cooper,J.B. (2007) XSTREAM: a practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. *BMC Bioinformatics*, **8**, 382–382.

Notredame,C. (2001) Mocca: semi-automatic method for domain hunting. *Bioinformatics*, **17**, 373–374.

Notredame,C. *et al*. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol*., **302**, 205–217.

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

Söding,J. and Lupas,A.N. (2003) More than the sum of their parts: on the evolution of proteins from peptides. *Bioessays*, **25**, 837–846.

Söding,J. *et al*. (2006) HHrep: de novo protein repeat detection and the origin of TIM barrels. *Nucl. Acids Res*., **34**, W137.

Sonnhammer,E.L. *et al*. (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucl. Acids Res*., **26**, 320–322.

Street,T.O. *et al*. (2006) The role of introns in repeat protein gene formation. *J. Mol. Biol*., **360**, 258–266.

Szklarczyk,R. and Heringa,J. (2004) Tracking repeats using significance and transitivity. *Bioinformatics*, **20**, i311.

Waterman,M.S. and Eggert,M. (1987) A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. *J. Mol. Biol*., **197**, 723–728.

Zhang,Y. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.