BIOINFORMATICS

Vol. 1 no. 1 2003
Pages 1–1

PILER: identification and classification of genomic repeats

Robert C. Edgar[1*] *and Eugene W. Myers*[2]

[1]*195 Roque Moraes Drive, Mill Valley, CA, U.S.A., bob@drive5.com.*
[2]*Department of Computer Science, University of California, Berkeley, CA, U.S.A.*

ABSTRACT

Repeated elements such as satellites and transposons are ubiquitous in eukaryotic genomes. *De novo* computational identification and classification of such elements is a challenging problem, so repeat annotation of sequenced genomes has historically largely relied on sequence similarity to hand-curated libraries of known repeat families. We present a new approach to *de novo* repeat annotation that exploits characteristic patterns of local alignments induced by certain classes of repeats. We describe PILER, a package of efficient search algorithms for identifying such patterns. Novel repeats found using PILER are reported for *H. sapiens*, *A. thalania* and *D. melanogaster*. The software is freely available at http://www.drive5.com /piler.

1 INTRODUCTION

Repetitive DNA is ubiquitous in eukaryotic genomes (Charlesworth *et al.*, 1994). Identification and classification of repeats is a fundamental annotation task that is important for several reasons. First, repeats are believed to play significant roles in genome evolution (e.g., (Bowen and Jordan, 2002)) and disease (e.g., (Buard and Jeffreys, 1997)). Second, mobile elements (transposons and retrotransposons) may contain coding regions that are hard to distinguish from other types of genes. Finally, repeats often induce many local alignments, complicating sequence assembly, comparisons between genomes and analysis of large-scale duplications and rearrangements..

Historically, repeat annotation has largely relied on similarity to hand-curated consensus sequences for known repeat classes. For example, RepeatMasker (Smit, A.F.A and Green, P., http://repeatmasker.genome.washington.edu) searches for nucleotide alignments to the RepBase database (Jurka, 2000). However, automated construction of such a repeat library "...remains a challenging bioinformatics problem" (Pevzner *et al.*, 2004). Previous programs addressing this problem are RepeatFinder (Volfovsky *et al.*, 2001), RECON (Bao and Eddy, 2002) and RepeatGluer (Pevzner *et al.*, 2004). All begin comparing the target DNA sequence or genome against itself to identify local alignments between different regions, but differ in the way they proceed to identify and classify repetitive elements.

Here we describe a novel approach to repeat identification and classification. In contrast to previous methods that attempt to explain *all* the off-diagonal local alignments or *hits*, we focus on identifying subsets of hits that form a pattern characteristic of a given type of repeat in such a way that the elements we do report are highly reliable. For example consider the following typical scenario: a segmental duplication occurs of a region containing several partial, degraded mobile elements of various types, followed by mutations of the duplicated regions including insertions (perhaps of additional mobile elements), deletions and rearrangements. Against this background of multiple, intertwined processes of genome evolution, existing methods may report many "elements" that are false, being in fact either fusions or fragments of the biologically true elements. Our strategy is to achieve high specificity at the sacrifice of some sensitivity.

Our terminology is as follows. An *element* is a sequence chosen to represent a set of similar genome subsequences produced by a single biological process. Informally, an element should correspond to a repeat library sequence. It may be constructed in various ways: for example as a centroid (i.e., the member that minimizes pair-wise distances to other members), the consensus derived from a multiple alignment, or from biological considerations, e.g. in the case of a mobile element as the copy believed most likely to be active. An *instance* of the element is a subsequence of the genome that aligns to all or a significant portion of the element. A *family* is a set of instances of a single element. An *intact instance* is one that can be globally aligned to the entirety of its element. Note that "intact" is not intended to imply biological activity; it could for example be a fossil transposon. An *isolated* instance is adjacent to unique genomic sequence at both termini. What is unique or globally alignable is defined by algorithm parameters that specify how much difference is allowed in an alignment and how much of the element must be aligned.

A *tandem array* (TA) is a contiguous series of intact instances of a single element. An element found in a tandem array is a *satellite*. A *dispersed family* (DF) has members that are typically separated in the genome, i.e. that are rarely

[*] To whom correspondence should be addressed.

or never found in tandem. Dispersed families are often mobile elements such as transposons or retrotransposons. A family that is not a TA but whose members cluster in the genome is a *pseudo-satellite* (PS). Pseudo-satellites are intermediate between satellites and dispersed repeats. In borderline cases the distinctions are arbitrary, but it is nevertheless useful to define a separate category. Biologically, satellites and pseudo-satellites may be associated with similar elements, and may not be distinguishable by low-resolution experimental techniques such as fluorescence *in situ* hybridization. For example, the SAR_DM element[1] in *D. melanogaster* is found both in satellites and pseudo-satellites. Algorithmically, limiting a search to neighboring regions has advantages that will be explained shortly. If the termini of an intact element align to each other, it has a *terminal repeat* (TR). TRs may be reverse complemented, as in Tc1 transposons, or not, as in Gypsy retrotransposons. TRs range in length from a few bases to a thousand or more. We use the term *random process* to mean an evolutionary mechanism that does not show strong preferences for particular sequence signals: the canonical example is segmental duplication. Unlike mobile elements and satellites, which tend to produce multiple copies of a single element, segmental duplications are unlikely to choose the same endpoints in two separate duplication events. Each of these classes of elements, TA, PS, DF, and TR, can produce distinctive signatures or patterns of hits that are highly to occur by chance, allowing reliable element constructions. Intact, isolated members of a dispersed family produce globally alignable regions (Figure 1). While *two* globally alignable regions are readily produced by segmental duplication, finding *three* or more strongly suggests a process that is specific to the sequence, such as a mobile element. Tandem arrays produce *pyramids* (Figure 2). Families with TRs produce overlapping hits that fall into two subsets: (a) those close to the main diagonal produced by a *single* instance, and (b) hits aligning TRs from opposite ends of *different* instances which are found at larger and varying distances form the main diagonal (Figure 3). In the following, we describe efficient search algorithms for these signatures that are implemented in our PILER package.

2 METHODS

2.1 Definitions

A hit is a local alignment of two regions Q and T, called its *images*. The endpoint coordinates of image Q are denoted by start(Q) and end(Q). Q = partner(T) is the *partner image* of T, and vice versa. We assume the length of the genome is L, the set of input hits is H, the number of hits is $N = |H|$, and the total length of all images in H is S.

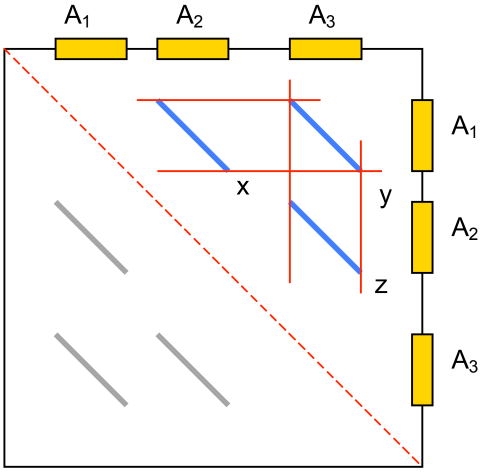

Fig 1. Signature induced by a dispersed family. This is a schematic view of a dot plot of a genome against itself with hits indicated by diagonal blue lines. Intact, isolated members of a dispersed family are seen as three globally alignable regions, inducing hits x, y and z of similar length. The dotted red line is the main diagonal of the dot plot. Gray alignments in the lower-left triangle are redundant by symmetry. Upper case letters represent elements; subscripts distinguish instances.

2.2 Finding local alignments

To find local alignments of minimum length (λ) and minimum identity (μ) we used the filtration method of (Rasmussen *et al.*, 2005) in a software tool PALS, Pairwise Alignment of Long Sequences, that we developed (Edgar, R.C. and Myers, E.W., http://www.drive5.com/pals). The indels required to make each local alignment explicit are not computed, reducing computational overhead. The algorithms described here require only hit endpoints, so each hit requires only four integers of storage. PALS is designed for repetitive sequences (for example, there is no limit on the total number of hits or on the number of hits that overlap a given region), and has optimizations for the special case of aligning a sequence to itself.

One such optimization is a *banded* search for alignments between regions separated by a maximum distance β (the *diameter* of the band), which allows sensitive searches in sequences of effectively unlimited length. For example, a whole-genome banded self-alignment of *A. thalania* (116 Mb) with λ=100, μ=0.9, β=10^5 requires five minutes on a 2 GHz desktop PC with 1 GB RAM.

2.3 Piles

A *pile* is a list of all hits covering a maximal contiguous region of copy count > 0. We construct piles as follows.

[1] We use RepBase identifiers for known elements.

```
Create vector $c$ of length $L$, set to zero.
Let $W$ be the set of all $2N$ images in $H$.
For each image $Q$ in $W$:
    for $x$ = start($Q$) to end($Q$):
        Set $c[x] = c[x] + 1$
{ Now $c[x]$ is the copy count of base $x$ }
Set $P = 0$ {$P$ is number of piles found so far}.
For $x = 1$ to $L$:
    if $c[x–1] = 0$ and $c[x] > 0$: {is $x$ start of new pile?}
        Set $P = P + 1$
    if $c[x] > 0$:
        Set $c[x] = P$
{ Now $c[x]$ is the identifier of the pile that covers base $x$,
  or zero if $x$ is unique }
Create $P$ empty piles. A pile is a list of images.
For each image $Q$ in $W$:
    Set $p = c$[start($Q$)]
    Add $Q$ to pile $p$.
```

Assuming a constant average hit length, this procedure is O(N) and is efficient in practice for typical input data. O(N log N) solutions that are not sensitive to hit length are also possible.

2.4 PILER-DF

PILER-DF is a search method designed to find intact, isolated members of a dispersed family (Figure 1). These are identified as sets of t or more globally alignable piles, where $t \geq 3$ is a parameter of the algorithm. (The minimum of three is required to distinguish a dispersed family from a segmental duplication). We define pile(Q) to be the pile containing Q, and *is-global-image*(Q) as true if Q covers a fraction $\geq g$ of the bases in pile(Q), where g <~ 1 is another parameter. Given a set of piles P the search is implemented in O(N) time and space as follows.

```
For each pile $p$ in $P$:
    For each image $Q$ in $p$:
        Let $T$ = partner($Q$)
        if is-global-image($Q$) and is-global-image($T$):
            Add edge $p$–pile($T$) to $G$
Find connected components of $G$ of order $\geq t$.
```

Each connected component is classified as a dispersed family and interpreted as a putative intact mobile element. Note the important difference between searching for globally alignable piles versus globally alignable hits. Let A_1, A_2 and A_3 be intact instances of element A, and a be a fragment of A. Hits a-A_1, a-A_2 and a-A_3 are globally alignable to each other, but a multiple alignment of these hits produces a consensus similar to a rather than the intact element A. By requiring piles, rather than hits, to align globally, this problem is avoided (because the intact copies align to each other and an a-A_i hit is therefore not globally alignable to the pile containing A_i). Note also that while many dispersed families are due to mobile elements, other processes can cause the DF signature. In particular, a functional region found in multiple segmental duplications (e.g., a paralogous exon) presumably tends to be more highly conserved than its surroundings, inducing globally alignable piles during a window in evolutionary time where the surrounding region is sufficiently degraded not to align but the functional region is still alignable. We call such features *buttes* (drawing an analogy between geological erosion and mutation).

2.5 PILER-PS

PILER-PS is a search for pseudo-satellites. The algorithm is essentially identical to PILER-DF, except that hits are identified by a banded search rather than a full search for self-alignments (see Section 2.2). Banded search offers several advantages for this task. It naturally enforces the defining requirement that pseudo-satellites are clustered. It also allows a faster and more sensitive search for hits, possibly enabling the discovery of more highly degraded instances. Finally it reduces noise produced by alignments to more distant, and perhaps more distantly related, repeats in other regions of the genome, which may result in cleaner piles and increased sensitivity.

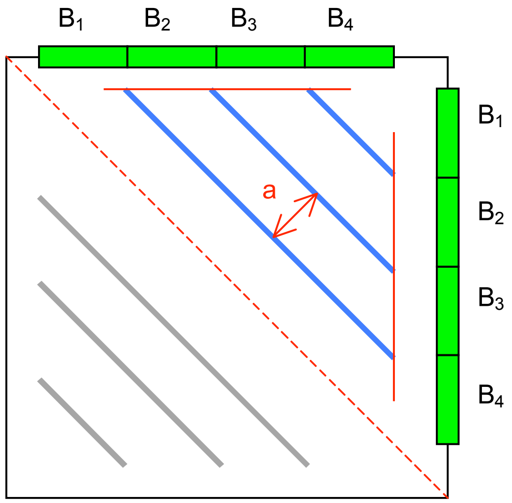

Fig 2. The signature induced by a tandem array. A tandem array appears in a dot plot as a triangle filled by diagonal stripes, a feature we call a pyramid. (To see a picture of a pyramid and its reflection, rotate the figure 45 degrees anticlockwise). Notation is similar to Fig 1. Hits are separated by diagonal distance a, which is the length of the repeated element (B). Hit endpoints align to each other, as suggested by the horizontal and vertical red lines.

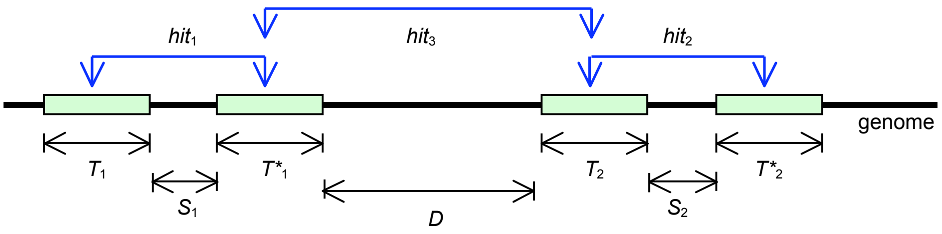

Fig 3. Terminal repeat search. First, candidate TRs are identified as hits such as hit_1 and hit_2 that align images of lengths (T) and separations (S) within bounds for mobile elements with terminal repeats. In a second pass, hits that align candidates to each other, such as hit_3, are identified by the following criteria: (1) its images are approximately a pair of candidate TRs (T^*_1, T_2) from different hits found in the first pass, (2) the candidates "joined" by this hit have similar separations (S_1, S_2), and (3) its images are sufficiently separated in the genome, i.e. have large enough D.

2.6 PILER-TA

PILER-TA is a search for tandem arrays, identified as pyramids (Figure 2). A pair of hits in a pyramid is recognized by (a) the alignment of their start and endpoints (horizontal and vertical red lines in Figure 2), and (b) requiring that the length of the shorter hit is at least half the length of the longer hit. These two criteria suffice to infer that the longer hit spans (possibly part of) a tandem array. To compare every pair of hits in H is $O(N^2)$, which may be prohibitively expensive. To reduce the search space, we observe that hits in a given pyramid necessarily belong to the same pile. By imposing a maximum length β on the array (more precisely, on the part of an array that can be identified by the search), we can further require that images are separated by a distance of at most β. We therefore use a banded search, allowing a fast and sensitive search for TAs in long sequences such as eukaryotic chromosomes. We define first(h) as the image in hit h with the smaller start coordinate, and last(h) as partner(first(h)). The search is implemented as follows.

```
For each pile $p$:
    Create an empty graph $G$
    For each pair of hits ($h$, $h'$) in $p$:
        Set $shorter\_length$ = min(|$h$|, |$h'$|)
        Set $longer\_length$ = max(|$h$|, |$h'$|)
        Set $Q$ = first($h$)
        Set T = last($h$)
        Set $Q'$ = first($h'$)
        Set $T'$ = last($h'$)
        Set $dS$ = (start($Q$) – start($Q'$)) / $shorter\_length$
        Set $dE$ = (end($T$) – end($T'$)) / $shorter\_length$
        if $shorter\_length$ / $longer\_length$ > 0.5 and
          |$dS$| < $m$ and |$dT$| < $m$:
            Add edge $h$–$h'$ to $G$
    Find connected components of $G$
```

Each connected component is classified as a tandem array. The parameter $0 \leq m << 1$ determines how closely the endpoints must align (red lines in Figure 2), expressed as a fraction of the length of the shorter hit. By default, m = 0.05. In a pyramid induced by an array of exact copies, the diagonal distances between hits are equal to the length of the repeated motif. We therefore estimate the element length e as the average distance between members of a cluster of closest pairs of diagonals in the pyramid (for robustness against missing hits due to alignment algorithm artifacts or marginal exclusion by the search criteria). Optimal boundaries between instances are harder to infer as any cyclic permutation of an estimated element will align almost equally well to the array. Even if optimal boundaries can be inferred for a single pyramid, the identification of a consensus sequence is best deferred until results from all satellite and pseudo-satellite searches can be clustered. We therefore employ simple heuristics to determine reasonable boundaries for a single pyramid. We find a subset of hits that align well (have small values of dS and dE), and are separated by diagonal distances in good agreement with the estimated element length. Boundaries are inferred from each such hit at a distance e from its endpoints, providing representative instances of the element as input to library construction (Section 2.8).

2.7 PILER-TR

PILER-TR is a search for families with terminal repeats, using the signature depicted in Figure 3. The first stage is a search for candidate TRs, which are identified as images within a length range typical of a terminal repeat (default: from 50 to 2000 bases) and separated by a distance reasonable for TR elements (default: from 50 to 15000 bases). Satellites and pseudo-satellites often induce many hits meeting these criteria, so these are identified and masked in a pre-processing step. A banded search is then performed for TR candidates. A second pass seeks hits that align candidate TR pairs to each other; this is done by a full search of the entire genome. Each such hit induces an edge in a graph connecting two candidates. Finally, connected components are identified and interpreted as putative families of TR elements. This is a straightforward application of the techniques described above; details are omitted for brevity.

2.8 Library construction and genome annotation

PILER search procedures are designed to find repeat elements with boundaries corresponding to individual biological events by finding instances that produce characteristic signatures. Typically only a small subset of the instances of an element are found in the patterns. For example, most instances of mobile elements in currently sequenced model organisms are either not intact or not isolated, and so are not directly reported by PILER-DF. More complete genome annotation therefore requires construction of a library of elements for use by a separate tool such as BLAST (Altschul *et al.*, 1990) or RepeatMasker. We use MUSCLE (Edgar, 2004a; Edgar, 2004b) to create multiple alignments of family members found by PILER, from which consensus or centroid sequences are produced. The library is screened for nucleotide and protein similarity to known repeats and known functional elements (e.g., genes), generating a report that is used for manual curation. Once this library has been curated, the genome can be scanned with BLAST or RepeatMasker to identify all intact and partial instances of the elements discovered by the PILER suite.

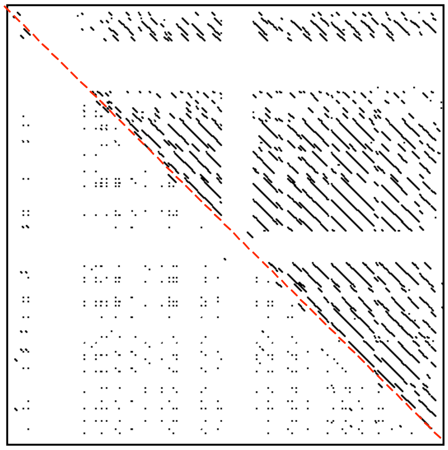

Fig 4. Hit plot for a 60kb region of *A. thalania* chromosome 1 (Genbank id 42592260, position 15050000) that is rich in satellites and pseudo-satellites. Each PALS hit is represented by a line connecting its endpoints, giving a picture that can be interpreted similarly to a dot plot. The upper-right triangle shows hits before masking. The lower-left triangle shows hits after masking by RepeatMasker or TRF, which independently give very similar results in this region. After masking by our PILER library, just one PALS hit remains (not shown).

3 RESULTS

In collaboration with several colleagues, we recently carried out a systematic validation of PILER-DF on three model organisms: *D. melanogaster*, *A. thalania* and *C. elegans*, finding many known mobile elements and a few that had not been previously reported (manuscript in preparation). We found that PILER-DF has high specificity, with sensitivity that varies depending on the genome. For example, in *D. melanogaster*, the PILER-DF library successfully reproduced 29 mobile element consensus sequences (with a successful match defined as > 90% mutual blastn coverage of a RepBase sequence). Most of these 29 sequences agreed on element boundaries to within 1% of the length of the matched RepBase sequence, and together account for 61% of the genome mobile element content, measured in bases annotated by RepeatMasker using the appropriate RepBase library. Seven other library sequences were either successful matches to terminal repeats or fragments of mobile elements. The remaining five sequences in the library had no significant blastn hit to RepBase. Two of these were due to paralogous exons; the other three were not readily identified. In *A. thalania*, a similar analysis found 33 successful matches to mobile elements, but these elements account for only 20% of the RepeatMasker mobile element annotation.

Here we describe some representative novel elements identified using PILER-TA, PILER-PS and PILER-TR in *H. sapiens* (Lander *et al.*, 2001; Venter *et al.*, 2001), *D. melanogaster* (Celniker *et al.*, 2002) and *A. thalania* (AGI, 2000). For examples, see Table 1.

3.1 Satellites and pseudo-satellites in *A. thalania*

To investigate the performance of PILER for *de novo* masking of satellites and pseudo-satellites, we performed the following experiment on the *A. thalania* genome, which is rich in these classes of repeat. As our current implementation of PILER is not designed to find short elements (< 50 bases), we first masked the sequence using Tandem Repeats Finder (TRF) (Benson, 1999), which locates tandem repeats of motifs from 1 to 500 bases. We then performed two search and masking steps. In each step, a library was constructed (Section 2.8) and masking performed by using blastn (e=10^{-5}) to identify a more complete set of instances. These steps used PILER-TA and PILER-PS, respectively. Hits were found using PALS parameters λ=100, μ=0.9, β=10^5. TA search identified 78 tandem arrays of elements ranging in length from 252 to 3413 bases. This demonstrates that PILER-TA can identify repeated motifs of length < 500 bases missed by TRF, and also those longer than the maximum allowed by TRF (by default 500 bases, or 1000 bases with the appropriate option selected). Using blastclust (http://www.ncbi.nlm.nih.gov/blast), we found that the 78 elements fall into 53 separate clusters and thus probably represent several distinct types. PILER-PS reported 191 pseudo-satellite families. We illustrate the improvement in masking achieved by the resulting PILER library on a 60kb region of chromosome 1. We find only one PALS hit after PILER masking, in contrast to TRF and RepeatMasker, which leave many visually obvious repeats unmasked (Figure 4). There is some circularity in this measure of success (because PALS generates the input to PILER), and in general the number of masked bases is a questionable measure of a repeat library as functional elements such as paralogs may be false-positive masked. However, we believe that in this case, the observed increase in masking is a strong indication of improved quality. See Table 1, entries (1) and (2) for examples of novel elements found in this analysis.

(#)	Seq	Acc	Type	Position	Length
(1)	A IV	30698542	TA	14348577	47991
(2)	A I	42592260	PS	8769822	6981
(3)	H 1	NT_004836.16	TA	10023940	894
(4)	H 1	NT_079497.2	PS	98834	1499
(5)	D	AABU01002684	TR	17722	7790
(6)	D	AABU01002755	TR	207940	4695

Table 1. Examples of novel elements identified by PILER searches. (#) is a number identifying the repeat, Seq is A I and A IV for *A. thalania* chromosomes I and IV respectively, H 1 for *H. sapiens* chromosome 1, D for *D. melanogaster* heterochromatin contigs; Acc is the Genbank accession number, Type is the PILER search method, Position is the sequence coordinate, and Length is the number of bases. For TA, the position and length are for the full array as reported by PILER-TA; for PS and TR the position and length locate an arbitrarily chosen member of the family. (1) A tandem array with motif length 2205 found in ~22 copies. The motif is partially masked by RepeatMasker (583 bases total in 6 hits). (2) A pseudo-satellite found in three well conserved instances (> 99% identical) by PILER-PS. This element has 37 hits to the full genome (blastn, e=10^{-9}) with at least one hit to every chromosome. It is not masked by RepeatMasker, and includes a coding region for unknown protein AAN12922.1. (3) A tandem array, motif length 88 found in ~10 copies. This motif is not masked by RepeatMasker. (4) A pseudo-satellite with 10 highly conserved instances identified by the PILER-PS search. It has a strong similarity to hypothetical protein DJ328E19.C1.1. (5) and (6) are dispersed families found in heterochromatin contigs of *D. melanogaster*, (BDGP Release 3). Both have strong similarity to Gypsy proteins. (6) is not masked by RepeatMasker, (5) is partially masked with several fragmentary annotations.

3.2 Satellites and pseudo-satellites in *H. sapiens*

In an attempt to identify novel satellites and pseudo-satellites in *H. sapiens w*e chose chromosome 1, NCBI build number 35, using the repeat masked (.mfa) sequence. We followed a similar procedure to Section 3.1, except that in this case we started from the downloaded masked sequence rather than using TRF. PILER-TA reported 67 tandem arrays of motifs ranging in length from 61 to 1950 bases. PILER-PS identified 55 pseudo-satellites of lengths ranging from 113 to 1518 bases. In some of these cases, we found that PILER misidentified the length of the repeat, reporting two or more concatenated instances of the true motif. Five of the pseudo-satellites found by PILER had partial matches to RepBase (blastn, e=10^{-5}), and 26 had significant hits to the NCBI non-redundant protein database (blastx, e=10^{-5}). Most protein hits were to hypothetical proteins, suggesting that some of these may be associated with uncharacterized mobile elements or satellite-associated repeats.

3.3 Gypsy LTRs in *D. melanogaster*

We ran PILER-TR on the *D. melanogaster* genome, including heterochromatin contigs, from BDGP Release 3. Masking was performed using RepeatMasker and TRF before running PILER-TR. 24 families were identified, several of which appear to be incompletely masked satellites or pseudo-satellites. Two families are well conserved and have strong protein hits to Gypsy proteins, suggesting that they are novel members of the Gypsy/LTR retrotransposon superfamily. See Table 1, entries (5) and (6).

DISCUSSION

Robust *de novo* computational identification and classification of genomic repeats is an important unsolved problem. The most obvious difficulties are caused by multiple interacting evolutionary processes. For example, most repeats due to mobile elements were presumably intact at the time they were inserted into the genome, but today are often found as fragmented, degraded copies that may be adjacent to repeats belonging to other families and / or embedded in segmental or tandem duplications. Functional regions within segmental duplications may be conserved, producing a repeat signature that can mimic a mobile element. The following problems must therefore be solved if a complete annotation is to be created from a set of local alignments: (1) boundaries between repeats of different families must be identified and distinguished from "random" repeats (segmental duplications), (2) repeat families must be clustered and full-length canonical sequences reconstructed where appropriate (e.g. for transposons), and (3) functional repeats (e.g. paralogs) must be distinguished from other types where needed for applications such as repeat masking. In this report we introduced a new approach to genomic repeat identification and classification based on searches for patterns of local alignments that are characteristic of particular classes of repeats. These classes are satellites, pseudo-satellites, mobile elements and mobile elements with terminal repeats, respectively. We also described PILER, a package of efficient algorithms that search for these signatures in a set of local alignments. We find that PILER searches have high specificity, with sensitivity that varies with the genome. Our discoveries of novel repeats show that PILER, followed by appropriate analysis and expert curation, enables improved annotation and repeat masking both in well studied model organisms and in new genomes. PILER searches typically require of the order of a second or less on a eukaryotic chromosome using current desktop computers.

It is natural to seek a comparison of PILER with other *de novo* repeat finding methods. PILER-TA is comparable with TRF as both are designed to find tandem repeats. The two methods have similar computational resource requirements and are easily able to process large genomes, such as mammals, on a desktop computer. While there is some overlap, we see the two methods as complementary given that TRF is better at finding arrays of short motifs while PILER-TA is able to find arrays of longer and more highly degraded motifs. While PILER-TA is reasonably successful in identifying the repeated motif in an array, further work is needed on this aspect of the algorithm to reduce errors such as misidentifying two or three concatenated instances of the true motif as the motif. Such errors are due to incomplete sets of overlapping hits from PALS. Other PILER search algorithms (DF, PS and TR) are not directly comparable with existing methods, including in particular RECON, RepeatFinder and RepeatGluer. These methods group repeats into related families, but, in contrast to PILER, do not explicitly attempt to identify boundaries induced by discrete evolutionary processes or to infer biological mechanisms such as mobile element insertions or segmental duplications. Instead, repeat boundaries are defined algorithmically and therefore cannot be compared unambiguously with a reference library such as RepBase or with methods using different definitions (arguably, such algorithms never make errors on their own terms).

While we have here followed convention and defined the goal of *de novo* repeat annotation in terms of library construction, we believe that more sophisticated approaches are needed. A functional region (or inactive sequence that is homologous to a functional region, such as a pseudogene) may be embedded inside a repeated region. Masking all significant local alignments to a repeat library may therefore mask functional sequence, fail to mask all self-similarities, or both. We believe that the goal should be a comprehensive annotation of both repetitive and functional regions, noting that these are not mutually exclusive. This could enable improved approaches to genome analysis: for example, alignment of two genomes might proceed as follows. Comprehensive repeat masking is performed prior to obtaining an initial set of local alignments, noting that this suppresses repeated functional regions. Such regions are treated as special cases by creating profiles from multiple alignments of each family of functional repeats in one genome; these are then used to search the other genome. This procedure avoids a combinatorial explosion of local alignments without discarding functional similarities.

The PILER software is freely available at http://www.drive5.com /piler.

4 ACKNOWLEDGEMENTS

The authors are grateful to Susan Celniker, Chris Smith and Roger Hoskins for many helpful discussions.

REFERENCES

AGI (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**(6814): 796-815.

Altschul, S.F., *et al*. (1990) Basic local alignment search tool. *J Mol Biol* **215**(3): 403-10.

Bao, Z. and Eddy, S.R. (2002) Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res* **12**(8): 1269-76.

Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**(2): 573-80.

Bowen, N.J. and Jordan, I.K. (2002) Transposable elements and the evolution of eukaryotic complexity. *Curr Issues Mol Biol* **4**(3): 65-76.

Celniker, S.E., *et al*. (2002) Finishing a whole-genome shotgun: release 3 of the Drosophila melanogaster euchromatic genome sequence. *Genome Biol* **3**(12): RESEARCH0079.

Charlesworth, B., Sniegowski, P. and Stephan, W. (1994) The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* **371**(6494): 215-20.

Edgar, R.C. (2004a) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**(1): 113.

Edgar, R.C. (2004b) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**(5): 1792-7.

Jurka, J. (2000) Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet* **16**(9): 418-20.

Lander, E.S., *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**(6822): 860-921.

Pevzner, P.A., Tang, H. and Tesler, G. (2004) De novo repeat classification and fragment assembly. *Genome Res* **14**(9): 1786-96.

Venter, J.C., *et al*. (2001) The sequence of the human genome. *Science* **291**(5507): 1304-51.

Volfovsky, N., Haas, B.J. and Salzberg, S.L. (2001) A clustering method for repeat analysis in DNA sequences. *Genome Biol* **2**(8): RESEARCH0027.