phRAIDER: Pattern-Hunter based Rapid Ad Initio Identification of Elemenrary Repeats

Carly E. Schaeffer[1], Nathniel D. Figueroa[1], Xiaolin Liu[2], and John E. Karro*[1,2,3,4]

[1] *Department of Computer Science and Software Engineering*
[2] *Cell, Molecular, and Structural Bology*
[3] *Department of Microbiology*
[4] *Department of Statisticcs, Miami University, Oxford, Ohio (USA)*

Abstract

This is the abstract.

Introduction

Transposable Elements (TEs) are genomic sequences thathad the capacity to insert copies of themselves into other genomic locations, resulting in homologous families of sequences spread across the genome. Present in almost every higher order genome (covering as much as 45% of the human genome and 90% of the maize genome [1, 2]), TEs have proved an important source of data in numerous studies of genomic structure (e.g. [3, 4, 5, 6]). But given their prevalence, it is important for those studying other aspect the genome to have TEs masked out – their bases replaced by N to allow for easy identification and filtering. Failure to filter can reek havoc with genomic analysis tools. For example, unfiltered TEs can trigger huge numbers of false positives in automated gene finding tool [7], as well as inflate tool runtime.

The best tools for repeat identification are RepeatMasker and nHMMER [8, 9], but both employ library-based search strategies using a pre-compiled description of sequences in the family (e.g. a ancestral sequence for BLASTing, or a profile HMM). But, much like we ask how the snow plow driver gets to work [10], we must ask how these libraries are compiled. Library-based tools are useless for the discovery of new families, and this are challanging to use on newly sequenced genomes.

Within mammalian species we can largely rely on homology relationships to port libraries across species. This does not hold so well in plants: in many cases TE composition of a given plant organism is species-specific. For example, a rice-based TE library will only identify 25% of the TEs in the maize genome [7].

To solve this problem we turn to *de novo* TE identification tools, identifying TEs using only the genome sequence information. A number of such tools are discussed in the literature. RECON uses WU-BLAST and PILER using LASTZ to compute self-alignments [11, 12, 13, 14]. RECON show good sensitivity but is computationally intensive and infeasible for use on whole genomes (requiring 60 hours for 18Mb rice genome in a 2013 study), while PILER achieves a good runtime with very low sensitivity . ReAS and RepeatScout [15] are based on k-mer searches, with the earlier showing less sensitivity than RECON [16, 15, 7]. RepeatGluer [17, 18]

*Corresponding Author

is based on a variation of DeBrujin graphs, which allows for a decomposition of TE families into domains, but is very, very computationally expensive. In Saha *et al.* the authors perform an extensive comparison of the tools, and conclude that RepeatScout is the best tool overall for assembled genomes, while ReAS the best when dealing with unassembled sequence fragments [19].

Elementary Repeats

Zheng and Lonardi approached the *de novo* identification problem using *elementary repeats* [20]. Similar to the RepeatGluer domains, elementary repeats are decompositions of TEs into basic building blocks. Identification of these building blocks are sufficient for the purpose of masking, and can be assembled into Transposable Elements for those interested in TEs themselves.

While it is notoriously difficult to mathematically model transposable elements [11], elementary repeats are more conducive to a formal description. For a given genome, a nucleotide sequence r is an elementary repeat if: (1) It is of at least length l (the length requirement); (2) there are at least f copies of r appear (the frequency requirement) (3) there is no proper substring of r of length $\leq l$ that appears in the genome independently of r (the minimality requirement); (4) r is a maximal string w.r.t (1-3) (the maximality requirement). Having proposed this definition, Zheng and Lonardi developed an identification algorithm that had a runtime quadratic in the query sequence size [20]. This was refined to linear time by He and also by Huo *et al.* [21, 22] based on variations of suffix tree approaches, but these appraoches are limited in their ability to handle sequence variation. As we are looking a genome size inputs, and TEs transposable elements naturally suffer from copy mistakes and accumulate instance-specific base substitutions over time, this is a significant limitation.

RAIDER

It was with the objective of creating a linear time identification algorithm that could handle variation through use of PatternHunter-like spaced seeds that we developed the prototype RIADER [23]. A rough implementation was first presented in Figueroa *et al.*, with more details in the Figueroa masters thesis [24, 25]. RAIDER was built along an alternate, but equivalent definition of elementary repeats based on l-mers (sequences of length l). Specifically, it was observed that the minimality condition could be rewritten as: There is no l-mer contained within elementary repeat r that appears in the genome more times than r. From there we make four core observations that form the basis for the RAIDER algorithm: (1) An l-mer cannot belong to two different elementary repeats; (2) Any l-mer in the genome that occurs f or more times is either an elementary repeat or belongs to one; (3) any two l-mers belonging to an elementary repeat must appear the same number of times in the genome; (4) If two sequences in the genome that are *maximally identical* (that is, cannot be extended in either direction and still be identical), these sequences cannot belong to a larger elementary repeat. (By "belong" we mean "is a substring of" – a definition we will be generalizing shortly.) For discussion and proof, see the Figueroa Thesis [25].

Based on these observations, we discover we can find all elementary repeats in a single scan of the genome. Specifically, as we scan from left to right, we track l-mer occurrences and identify multiple copies of the same l-mer. When we find the same sequence of l-mers occurring multiple times in a row, we can mark it as a tentative family, then break it down later if we discover violations of the minimality condition. The algorithm is summarized in Figueroa *et al.*, and discussed in detail in the Figueroa Thesis [24, 25].

Results of the preliminary implementation were promising. On human chromosome 22 we saw a 12× speedup over RepeatScout to RAIDER (2344 seconds to 192 seconds), while coverage of the RepBase [26] ancestral sequence improved (77% to 84%), while on mouse chromosome 19 we saw the same speedup with a significant drop in coverage (53% to 30%). On the full human genome RAIDER ran in 6.3 hours, while RepeatScout was unable to complete it run. For details, see Figueroa *et al.* [24].

Spaced Seeds

PatternHunter, a very successful augmentation to BLAST [23, 27], is based on the notion of *spaced seeds*: improving the sensitivity of string matching based algorthms by allowing wild-cards in the match. That is, instead of requiring two strings matching in 12 consecutive characters, we might instead require two six-character exact matches seperated by one base which may or may not match (represented by the *seed pattern* 11111101111111), or perhaps three consecutive four-character exact matches sperated by two bases each (1111001111001111). It has been demonstrated that certain seed patterns can indunce significant imporvements in BLAST sensitivity with out time penalty, though what makes a good pattern is not well understood.

RAIDER was designed with the intent of employing the spaced seed strategy, but this was only implemented heuristically for the Figuera *et al.* paper [25] – serving primarily as a proof-of-concept. Since its release we have develoed a formal model of elementary repeats that incorporated spaced seeds, and from that developed phRAIDER (PatternHunter-based RAIDER). phRAIDER is a fast tool for the identification and making of transposible elements in both assembled and unassembled genomes aht outpreforms RptScout and other establihsed tools. Code is free available under the Gnu GPL lisence (v. 3) and may be obtained [NEED WEB ADDRESS].

In the following we will present a new theoretical model for *seeded elementary repeats*, followed by an algorithm for identifying them. In the results section we will show the result of using that algorithm for masking transp

In our Methods section we present the new model that allows us to extend RAIDER to correctly use spaced seeds, and briefly outline the algorithm (with more details provided in the supplementary materials), with a analysis of phRAIDER performance in our Results.

Model

Our goal is to redefine the concept of transposbile elements to accomodate a spaced seed strategy. In the next section we will describe our identification algorithm, and then quantified its success in masking transposable elements. But we will start here with a brief outline of our theoretical model (with a more detailed description in the appendix).

We first need to redfine out terminimology regarding elementary repeats. Under the Z&L definition, all instances of one elementary repeat have the exact same sequence, and hence can be descrbied by a single string. As we are allowing for variation, we will describe out instance set with a *sequence descriptor* consisting of bases letters and the wild-card character * (e.g. AAC*GG would describe a set of sequences with starting with an AAC, ending with a GG, and having any character in between). Given a binary string s representing a spaced seed, we say a sequence descriptor r is *consistent* with s if we can align s to a substring of r such that every $*$ in r aligns with a 0 character in s. (Hence $ACG**T*A$ is consistent with the seed 11001, but not 11011.)

Given a sequence descriptor r, we can *decompose* r with respect to s by taking every length $|s|$ substring of r that is consistant with s, replacing all letters of r that match to a 0 in s with a $*$, and creating a set from the results. (Hence for $s = 11011$ and $r = AA*CCGTT$, the decomposition would be $\{AA*CC, CC*TT\}$.) We say s *covers* r is every base in r is contained in at least on string in the decompositio. (Hence the previous s does not cover $AAA*CC$, as the first base in not in any of the strings of the decomposition.)

We can now modif the previous definition of elementary repeats as follows:

Definition 1 *Given a genomic sequence G, an integer f, and a spaced seed s, a sequence descriptor r describes an elementary repeat if it meets the four (moified) requirements of an elementary repeat:*

- *Structure requriement: s covers r.*
- *Frequency requirement: There are at least f substrings of G that match r.*

- *Minimality requirement: For every string t in the decomposition of r w.r.t s, the number of occurences of t in G is equal to the number of occurneces of r in the genome.*
- *Maximality requirement: There is no sequence descriptor r' of r that contains r as a proper substring and satisfies conditions 1-1.*

Theorem 1 *When the seed s has no 0 characters, this definition of elementary repeats is equivilent to the Z&L definition.*

phRAIDER Algorithm

References

[1] Lander, E.S.e.a.: Initial sequencing and analysis of the human genome. Nature **409**(6822), 860–921 (2001)

[2] SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z., Bennetzen, J.L.: Nested retrotransposons in the intergenic regions of the maize genome. Science (New York, NY) **274**(5288), 765–768 (1996)

[3] Arndt, P.F., Hwa, T.: Identification and measurement of neighbor-dependent nucleotide substitution processes. Bioinformatics (Oxford, England) **21**(10), 2322–2328 (2005)

[4] Karro, J.E., Peifer, M., Hardison, R.C., Kollmann, M., von Grünberg, H.H.: Exponential decay of GC content detected by strand-symmetric substitution rates influences the evolution of isochore structure. Molecular Biology and Evolution **25**(2), 362–374 (2008)

[5] Mugal, C.F., von Grünberg, H.H., Peifer, M.: Transcription-Induced Mutational Strand Bias and Its Effect on Substitution Rates in Human Genes. Molecular Biology and Evolution **26**(1), 131–142 (2008)

[6] Hardison, R.C.: Covariation in Frequencies of Substitution, Deletion, Transposition, and Recombination During Eutherian Evolution. Genome research **13**(1), 13–26 (2003)

[7] Jiang, N.: Overview of repeat annotation and de novo repeat identification. Methods in molecular biology (Clifton, N.J.) **1057**(Chapter 20), 275–287 (2013)

[8] Smit, A.F., Hubley, R., Green, P.: RepeatMasker Open 4.0

[9] Wheeler, T.J., Eddy, S.R.: nhmmer: DNA homology search with profile HMMs. Bioinformatics (Oxford, England) **29**(19), 2487–2489 (2013)

[10] Pratchett, T.: Hogfather: A Novel of Discworld. HarperPrism, New York

[11] Bao, Z., Eddy, S.R.: Automated de novo identification of repeat sequence families in sequenced genomes. Genome research **12**(8), 1269–1276 (2002)

[12] Edgar, R.C., Myers, E.W.: PILER: identification and classification of genomic repeats. Bioinformatics (Oxford, England) **21 Suppl 1**(Suppl 1), 152–8 (2005)

[13] Lopez, R., Silventoinen, V., Robinson, S., Kibria, A., Gish, W.: WU-Blast2 server at the European Bioinformatics Institute. Nucleic acids research **31**(13), 3795–3798 (2003)

[14] Harris, R.S.: Improved pairwise alignment of genomic DNA. PhD thesis, The Pennsylvania State University (2007)

[15] Price, A.L., Jones, N.C., Pevzner, P.A.: De novo identification of repeat families in large genomes. Bioinformatics (Oxford, England) **21 Suppl 1**(Suppl 1), 351–8 (2005)

[16] Li, R., Ye, J., Li, S., Wang, J., Han, Y., Ye, C., Wang, J., Yang, H., Yu, J., Wong, G.K.-S., Wang, J.: ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. PLoS computational biology **1**(4), 43 (2005)

```
Algorithm phRAIDER
  function SAMEOFFSET($l$-mer $v_1$, $l$-mer $v_2$)
      # Is the distance between the first two occurences of $v$ equal to the
      # the distance betwen the most recent two occurences?
      return $H[v_2][0] - H[v_1][0] == H[v_2][-1] - H[v_1][-1]$

  function PHRAIDER(Genome $G$, Seed $s$, MinFrequence $f$)
      # We are going to maintain a $Q$ of up to $|s|$ families, representing the families
      # of the last $|L|$ $l$-mers seen, sorted by the position of those last $l$-mers.
      Queue<Family> $Q$;

      # Iterate over all $l$-mer positions in $G$
      for $i \leftarrow 1 \rightarrow |G| - |s|$ do
          # Remove any families whose last $l$-mer did not occur in the last $i$ bases.
          while $i-$location($Q$, finish, last, end)$> |s|$ do
              $Q$.dequeue()

          # Get the $l$-mer occuring at $G_i$ and remove letters corresponding to 0s in $s$.
          $v \leftarrow \text{seeded}(G[i], x)$

          # Add to list of $v$'s positions.
          $H[v].push(i)$

          if $|H|(v)| == 2$ then
              # $v$ either is the beginning of a family, or should be combined with a
              # previous $l$-mer to form a family.
              $F = \text{arg\_find}_{F \in Q}\{\text{sameOffset}(F.\text{start}(), v)\}$
              if $F$ then
                  $Q$.requeue($F$)
              last($Q$).addLmer($v$)
          else if $|H[v]| > 2$ then
              # Is $v$ the proper continuation of a family, or does it need to be split off?
              $F \leftarrow v$.family()
              if $H[F.\text{lastSeen}(i)][-1] < H[Q.\text{front}().\text{lastSeen}()]$ then
                  # $v$ belongs to a family whose last instance is too far back.
                  if $v \neq F.\text{first}()$ then
                      # $v$ is not the first $l$-mer of its family
                      $F' \leftarrow F.\text{split}(F.v)$
                  else
                      $F' \leftarrow F.\text{split}(F.\text{lastSeen}())$
                  $Q$.enqueue($F'$)
              else
                  if $v == F.\text{first}()$ then
                      # $v$ is the first $l$-mer of $F$, so $F$ is moved to the back of $Q$
                      $Q$.requeue($F$)
                  else if sameOffSet($F$.first(), $v$) then
                      $F$.setLastSeen($v$)

      # Final break up of families as needed; filter families with frequency $\leq f$.
      tieLooseEnds()
```

[17] Pevzner, P.A., Pevzner, P.A., Tang, H., Tesler, G.: De novo repeat classification and fragment assembly. Genome research **14**(9), 1786–1796 (2004)

[18] Zhi, D., Raphael, B.J., Price, A.L., Tang, H., Pevzner, P.A.: Identifying repeat domains in large genomes. Genome biology **7**(1), 7 (2006)

[19] Saha, S., Bridges, S., Magbanua, Z.V., Peterson, D.G.: Empirical comparison of ab initio repeat finding programs. Nucleic acids research **36**(7), 2284–2294 (2008)

[20] Zheng, J., Lonardi, S.: Discovery of Repetitive Patterns in DNA with Accurate Boundaries. In: Fifth IEEE Symposium on Bioinformatics and Bioengineering (BIBE'05), pp. 105–112. IEEE, ??? (2005)

[21] He, D.: Using suffix tree to discover complex repetitive patterns in DNA sequences. Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference **1**, 3474–3477 (2006)

[22] Huo, H., Wang, X., Stojkovic, V.: An Adaptive Suffix Tree Based Algorithm for Repeats Recognition in a DNA Sequence. In: 2009 International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing, pp. 181–184. IEEE, ??? (2009)

[23] Li, M., Ma, B., Kisman, D., Tromp, J.: Patternhunter II: highly sensitive and fast homology search. Journal of bioinformatics and computational biology **2**(3), 417–439 (2004)

[24] Figueroa, N.: RAIDER: Rapid Ab Initio Detection of Elementary Repeats. PhD thesis, Oxford (2013)

[25] Figueroa, N., Liu, X., Wang, J., Karro, J.: RAIDER: Rapid Ab Initio Detection of Elementary Repeats. In: Advances in Bioinformatics and Computational Biology, pp. 170–180. Springer, Cham (2013)

[26] Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J.: Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and genome research **110**(1-4), 462–467 (2005)

[27] Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research **25**(17), 3389 (1997)