Genomics are far from being random but are made up of systematically ordered and information rich patterns. These repeated sequence patterns have been vastly utilized for their fundamental importance in understanding the genome function and organization. To this end, a comprehensive toolkit, RepEx, has been developed which extracts repeat (inverted, everted and mirror) patterns from the given genome sequence(s) without any constraints. The toolkit can also be used to fetch the inverted repeats present in the protein sequence(s). Further, it is capable of extracting exact and degenerate repeats with a user defined spacer intervals. It is remarkably more precise and sensitive when compared to the existing tools. An example with comprehensive case studies and a performance evaluation of the proposed toolkit has been presented to authenticate its efficiency and accuracy.

© 2013 Elsevier Inc. All rights reserved.
1.1 Exact repeat

\[ 5\-'ATGCAGTGC\text{-}CGC\text{-}CATA\text{-}ACCA\text{-}CAT\text{-}CA\text{-}CAGTG\text{-}TCGA\text{-}3\- \]
\[ 3\-'TACG\text{-}TCACGGCGGT\text{-}ATTGGT\text{-}AGTG\text{-}TCACGACT\text{-}5\- \]

1.2 Degenerate repeat

\[ 5\-'ATGCAGTGC\text{-}CGCC\text{-}ACCAT\text{-}CA\text{-}CAGTG\text{-}TCGA\text{-}3\- \]
\[ 3\-'TACG\text{-}TCACGGCGGT\text{-}ATTGGT\text{-}AGTG\text{-}TCACGACT\text{-}5\- \]

Fig. 1. Schematic representation of the exact and degenerate repeats. The arrows correspond to the organization and orientation of the repeats. (1.1) The figure depicts the exact direct repeat where the pattern is identical to the repeating unit i.e. a perfect repeating DNA fragment. (1.2) The figure depicts the degenerated direct repeat where the repeating units are not identical but are 71.4% similar.

level can be time consuming and computationally exhaustive. Thus, there is a paramount need to identify these genomic repeats in a fast and efficient way. To this end, RepEx, an efficacious toolkit has been proposed which identifies the above mentioned repeats from multiple genome/protein sequences in a near linear time. In general, methods for identifying repeats can be developed for finding exact and degenerate repeats. RepEx is capable of finding both. To extract direct and inverted repeats, there are many algorithms available [12-16]. However, for detection of everted and mirror repeats, to the best knowledge of the authors, there exists no efficient and robust toolkit. Thus, the core of the proposed toolkit has been specifically designed and optimized for the same. The proposed toolkit utilizes suffix tree to find multiple local alignments which in turn is employed to identify the repeats in genome sequence(s).

2. Methods

The core of this toolkit (Fig. 3) involves finding multiple local pairwise alignments between the query and its ancillary sequence (it is the modified query sequence based on the user defined repeat options). Then appropriate logical filters are applied to recognize the repeats among the generated local pairwise alignments. The operations are summarized in the following steps:

a) Generates necessary ancillary sequence(s) for a user given input and repeat options.

b) Perform pairwise alignments to find all possible multiple local alignments between the input and its ancillary.

c) Identify repeats from the resulting multiple local alignments.

The pairwise alignments are performed by searching through the suffix tree generated for the given sequence. This method facilitates a compact representation by storing all possible suffixes for the given query strings. In order to govern the local alignments through suffix tree, an open source package, MUMmer [17] is deployed. After the local alignments are determined, the repeats are identified through logical filters as described in Section 3. Locally developed PERL scripts are utilized for this purpose. In the proposed toolkit, RepEx, the use of suffix tree has provided the advantage of both near linear time and space complexity. Therefore, for a large genome, the running time increases linearly with size making it efficient, faster and more robust.

3. Implementation

RepEx can be used to identify the repeats from both genome and protein sequence(s). Let \( S \) be a sequence, \( N \) be the length of the sequence \( S \), and \( S, \bar{S} \) and \( \bar{S} \) denote the complement, reverse and reverse complement of \( S \), respectively. Circumflex \( \bar{S} \), in this context, symbolizes an operation for customary base complementarity e.g., if \( S = ATGC \) then \( \bar{S} = TACG \). Caron \( S \) (\( \bar{S} \)) symbolizes the operation for reverse complementarity, so \( \bar{S} = GCAT \) and \( \bar{S} \) symbolizes reverse, which corresponds to CGTA. Let \( S\_\text{S}[i][j] \) and \( S\_\text{S}[p+1] \) represent the local alignment between the two sequences \( S_1 \) and \( S_2 \), where \( i, j, p \) are the positions of the subsequence in \( S_1 \) and \( S_2 \), respectively. The subsequent subsections describe various conditions deployed in the toolkit to identify inverted, mirror and everted repeats.

3.1. Inverted repeat

Let \( i, j \) and \( p, q \) be the respective positions of the subsequence of \( S \) and \( \bar{S} \). For a local alignment to be identified as an inverted repeat, it has to satisfy the following logical condition.

\[ (((N-q+1)\leq i)\&\&((N-p+1)\leq i)) \]

Making \( S[N-p+1]..S[N-q+1] \) an inverted repeat of \( S[i]..S[j] \) in the sequence \( S \).
3.2. Mirror repeat

A local alignment is recognized as a mirror repeat, if it satisfies the following condition, where $i$, $j$, $p$, and $q$ are the respective positions of the subsequence of $S$ and $\hat{S}$.

\[
((N-q+1)<i) \&\&(N-p+1)<i)
\]

Making $S[N-p+1]..S[N-q+1]$ a mirror repeat of $S[i]..S[j]$ in the sequence $S$.

3.3. Everted repeat

Let $i$, $j$, $p$, and $q$ be the respective positions on the substrings of $S$ and $\hat{S}$. For a local alignment to be identified as an everted repeat, it has to satisfy the following condition.

\[
(p<i) \&\&(q<i)
\]

Making $S[p]..S[q]$ an everted repeat of $S[i]..S[j]$ in the sequence $S$. These conditional filters facilitate the identification of the exact repeats from the given sequence(s). To cull out the degenerate repeats, RepEx utilizes the recognized exact repeats as seeds and links them by allowing gaps and mismatches.

4. Options

RepEx uses four nucleobases for DNA and 20 standard amino acid residues for proteins. Users can provide the minimum length of the repeats to be identified and can also decide upon the proximity of the repeating copies to be as local, global or can define their own spacer intervals. If the number of nucleobases between the two repeat copies in a given sequence is less than 100, then these copies are treated as local repeats. On the other hand, if the number of nucleobases exceeds 100, they are treated as global repeats [18]. RepEx provides an option for the users to extract either exact or degenerate repeats or both together with user defined spacer intervals.

5. Case study

5.1. Inverted repeat in DNA

Many a times, bacteria are faced with environmental challenges which when answered appropriately correspond to its survival. They
respond to the relative changes in nutrient concentrations with the course of time. Complex regulatory mechanisms have evolved to coordinate their cellular requirements. Carbon catabolite repression is such an event which controls the regulatory interactions in various microorganisms. Lately, catabolite repression has been extensively studied in *Escherichia coli* [19]. The catabolite repression is facilitated by catabolite repressor/activator (Cra) proteins, which regulates the carbohydrate metabolism. Cra, a member of LacI–GalR family, binds to an inverted repeat present in the promoters of the target operons and activates/represses the transcriptional process. The complete genome of *E. coli* (GenBank accession No. U00096; length: 4639675 base pairs) was given as input to the proposed toolkit to extract both exact and degenerate inverted repeats [19] of a minimum length of 5 nucleobases with no spacers. The results were obtained in 14 s (performed on a 64 bit Linux box with 3.0 GHz Intel Core 2 Duo processor and 8 GB RAM). As anticipated, RepEx identified the inverted repeat TGAAAC*GTTTCA* [query position from 2261606 to 2261617] (where * indicates the center of symmetry) to which Cra binds, causing either activation or inhibition of transcription.

5.2. Mirror repeats in DNA

Mirror repeats are known to form triplex H-DNA; these repeats are usually purine or pyrimidine rich sequences. It is demonstrated that the structures of H-DNA are intrinsically mutagenic in mammalian cells [20]. They are known to induce large scale mutations such as deletions and/or rearrangements at a higher frequency [21]. Identification of H-DNA forming mirror repeats helps to pinpoint sequences that are susceptible to high frequency mutations, which in turn helps in understanding the fundamental molecular basis of diseases like autosomal dominant polycystic kidney disease (ADPKD). Mutations in the human *PKD1* (polycystic kidney disease) gene account for up to 85% of ADPKD cases [22]. The gene *PKD1* is shown to consist of several H-DNA forming mirror repeats [23]. RepEx was used to examine the mirror repeats in *PKD1* gene of human chromosome 16 (GenBank accession No. NC_000016.9; length: 90354753 base pairs). Mirror repeats with spacer length less than 1000 base pairs are extracted for both exact and degenerate repeats. A total of 24 degenerated mirror repeats with repeat lengths ranging from 30 to 324 base pairs and seven exact mirror repeats were identified by RepEx in 2.5 min (performed on a 64 bit Linux box with 3.0 GHz Intel Core 2 Duo processor and 8 GB RAM). From these results (Table 1), it is evident that the repeat sequences are purine or pyrimidine rich, indicating a high chance of H-DNA formation.

![](image)

**Table 1**
Table displaying exact and degenerated mirror repeats in *PKD1* gene of the human chromosome 16.

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Query position</th>
<th>Repeat copy</th>
<th>Query position</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCCCTCCCTCCCTCCCT</td>
<td>30854..30869</td>
<td>TCCCCCTCCCTCCCT</td>
<td>31505..31520</td>
</tr>
<tr>
<td>CCCCTCCCTCCCTCCCT</td>
<td>30882..30899</td>
<td>TCCCCCTCCCTCCCT</td>
<td>30959..30976</td>
</tr>
<tr>
<td>CCCCTCCCTCCCTCCCT</td>
<td>30882..30899</td>
<td>TCCCCCTCCCTCCCT</td>
<td>30911..30928</td>
</tr>
<tr>
<td>CAAAAAAAAAAAAAAAAA</td>
<td>6330..6348</td>
<td>AAAAAAAAAAAAAAAAAAC</td>
<td>6719..6737</td>
</tr>
<tr>
<td>CAAAAAAAAAAAAAAAAAT</td>
<td>6340..6358</td>
<td>TAAAAAAAAAAAAAAAAAA</td>
<td>6718..6736</td>
</tr>
<tr>
<td>GTTTTTTTTTTTTTTT</td>
<td>2793..2808</td>
<td>TTTTTTTTTTTTTTTG</td>
<td>3308..3325</td>
</tr>
<tr>
<td>GTTTTTTTTTTTTTTT</td>
<td>2793..2810</td>
<td>TTTTTTTTTTTTTTTT</td>
<td>3028..3045</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Query position</th>
<th>Repeat copy</th>
<th>Query position</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCCCTCCCTTTTTTTTTT</td>
<td>2793..2810</td>
<td>TTTTTTTTTTTTTTTT</td>
<td>3308..3325</td>
</tr>
</tbody>
</table>

**Table 2**
Table displaying exact inverted repeats in human cellular tumor antigen p53 isoform A.

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Query position</th>
<th>Repeat copy</th>
<th>Query position</th>
</tr>
</thead>
<tbody>
<tr>
<td>APAPAAP</td>
<td>74..80</td>
<td>PAAPAAPA</td>
<td>82..88</td>
</tr>
<tr>
<td>LTI</td>
<td>252..254</td>
<td>ITL</td>
<td>255..257</td>
</tr>
</tbody>
</table>

**Table 3**
Table displaying exact inverted repeats in human cellular tumor antigen p53 isoform A.
Inverted repeats (proteins and DNA), mirror repeats (DNA) and everted repeats (DNA) were manually inserted into the random sequences. RepEx was employed to identify repeats (inverted, mirror and everted repeats) of length greater than or equal to 8. The proposed tool identified all the manually inserted repeats (i.e. true positive hits) without producing any false positives (Table 3).

5.5. Performance evaluation

The performance of RepEx is evaluated with that of the existing inverted repeat detection tools IRF [15] and einverted [16]. Four different genome sequences of varying lengths are given as input to each of the above mentioned tools. All computations (i.e. the extraction of exact and degenerated inverted repeats) are performed with default parameters of the individual tools on a 64 bit Linux box with 3.0 GHz Intel Core 2 Duo processor with 8 GB RAM. The computation time utilized by RepEx is significantly less when compared to that of IRF and einverted (Table 4). Besides the computation time, a differential delineation of the results is performed. The results are first cataloged into their respective classes (exact and degenerate) and a PERL script is executed to reckon the results (Table 5). From this, it is apparent that RepEx is remarkable in culling exact repeats and is relatively better in extracting degenerate repeats. Therefore, RepEx is highly suitable for detecting exact and degenerate repeats in large genome sequences.

5.6. Sensitivity and precision of RepEx

To evaluate the precision and sensitivity of RepEx, its repeat extracting competency is compared with the above mentioned tools. To facilitate this evaluation, two simulated test sequences are designed — with (positive test sequence) and without (negative test sequence) inverted repeat. For the positive test sequence, a set of inverted repeats of length seven is strategically inserted in the simulated random DNA sequence to measure the true positives. To measure the false positives, a negative test sequence is generated by excluding the incorporated inverted repeats from the positive test sequence. All the tools considered in this study are allowed to extract the inverted repeats (of length seven) from the positive and negative test sequences. Based on the results, a confusion matrix is constructed to calculate the precision and sensitivity (also known as recall) of individual tools. Both the precision and the sensitivity ranges over 0 to 1 and the value one corresponds to the perfect extraction of repeats. It is observed that, the precision of both RepEx and IRF is maximum (i.e., precision is equal to one), which implies that the tools (RepEx and IRF) have the maximum positive prediction value as they extracted most pertinent (i.e. the extracted repeats) from the positive and negative test sequences, thus, making both its precision and sensitivity the least (i.e., zero).

As precision and sensitivity are two independent measures addressing different attributes of the individual tool's capability, a single number evaluation metric, F1 score (also known as F score) is implemented to measure the overall competency of the individual tools. It is a standard statistic measure that compromises between sensitivity and precision and is the harmonic mean of precision and sensitivity. F1 score ranges over 0 to 1. The F1 score for RepEx is 1, while for IRF it is 0.28. Thus, from the above three measures, it is evident that the proposed toolkit, RepEx has the highest true positive rate and has the ability to extract virtually all of the relevant repeats from the subjected query sequence.

### Table 3

Table displaying exact repeats from the simulated protein and DNA sequences.

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Query position</th>
<th>Repeat copy</th>
<th>Query position</th>
</tr>
</thead>
<tbody>
<tr>
<td>From</td>
<td>To</td>
<td>From</td>
<td>To</td>
</tr>
<tr>
<td>3.1 Inverted repeats in simulated protein sequence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LDDQALER</td>
<td>545</td>
<td>553</td>
<td>RELAEILDL</td>
</tr>
<tr>
<td>LFENIQILILAP</td>
<td>125</td>
<td>137</td>
<td>PAULUSQUNEIL</td>
</tr>
</tbody>
</table>

### Table 4

Evaluation results using various tools (RepEx against IRF and einverted).

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Size (MB)</th>
<th>Length (mbp)</th>
<th>RepEx</th>
<th>IRF</th>
<th>Einverted</th>
</tr>
</thead>
<tbody>
<tr>
<td>S. cerevisiae (Chr-4)</td>
<td>1.5</td>
<td>1.5</td>
<td>1.7</td>
<td>12.94</td>
<td>15.08</td>
</tr>
<tr>
<td>M. tuberculosis</td>
<td>4.3</td>
<td>4.4</td>
<td>8.31</td>
<td>66.51</td>
<td>41.90</td>
</tr>
<tr>
<td>A. lyra (Chr-5)</td>
<td>20.6</td>
<td>21.2</td>
<td>35.19</td>
<td>270.69</td>
<td>203.37</td>
</tr>
<tr>
<td>H. sapiens (Chr-11)</td>
<td>133</td>
<td>135</td>
<td>331.49</td>
<td>946.13</td>
<td>1217.25</td>
</tr>
</tbody>
</table>

a MB — Mega Bytes.
b mbp — Million base pairs.
6. Conclusion

A fast and robust toolkit, RepEx, is proposed to identify the major repeat types available in the genome and protein sequences. Also, it detects both exact and degenerate repeats with a user defined space interval. Based on the precision, sensitivity and F1 score, it is evident that RepEx proves to be accurate compared to the existing tools. Further, it works for multiple gene or protein sequences. This is the first toolkit which is capable of extracting the above discussed repeats from multiple and large genome sequences in a near linear time.

Acknowledgments

The authors acknowledge the facilities offered by the Interactive Graphics Facility and Supercomputer Education and Research Centre. One of the authors (KS) thanks the Department of Information Technology (DIT), Government of India for financial support.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2013.07.005.

References