

The Repeat Pattern Toolkit (RPT): Analyzing the Structure and Evolution of the *C. elegans* Genome

Pankaj Agarwal and David J. States

Institute for Biomedical Computing
Washington University
700 S. Euclid Avenue
St. Louis, MO 63110
{agarwal,states}@ibc.wustl.edu

Abstract

Over 3.6 million bases of DNA sequence from chromosome III of the *C. elegans* have been determined. The availability of this extended region of contiguous sequence has allowed us to analyze the nature and prevalence of repetitive sequences in the genome of a eukaryotic organism with a high gene density.

We have assembled a Repeat Pattern Toolkit (RPT) to analyze the patterns of repeats occurring in DNA. The tools include identifying significant local alignments (utilizing both two-way and three-way alignments), dividing the set of alignments into connected components (signifying repeat families), computing evolutionary distance between repeat family members, constructing minimum spanning trees from the connected components, and visualizing the evolution of the repeat families.

Over 7000 families of repetitive sequences were identified. The size of the families ranged from isolated pairs to over 1600 segments of similar sequence. Approximately 12.3% of the analyzed sequence participates in a repeat element.

Introduction

The genomes of humans and other higher organisms contain many sequences that are repeated one or more times. These repetitive elements range from mononucleotide tracts (for example, poly(A) repeats) to large complex segments (tens of kilobases), and from exact duplicates to highly-mutated copies (just detectable). The human genome is well known for the variety and number of repeated elements. In fact, the most prevalent repeat in the human genome—*Alu*—is about 300 base pairs long and represents about 5–10% of all the DNA in the genome (Deininger & Schmid 1979; Britten *et al.* 1988; Jurka, Walichewicz, & Milosavljevic 1992). *Alu* elements appear to be associated with transcriptionally active genes.

^oCitation: *Proceedings Second International Conference on Intelligent Systems for Molecular Biology* (editors R. Altman, D. Brutlag, P. Karp, R. Lathrop, and D. Searls) AAAI press, 1994, pages 1–9.

Some repetitive sequence elements, such as the *Line-1* family in humans, appear to be derived from retroviruses. Some copies of both *Alu* and *Line-1* are still functionally active, producing both RNA and protein products (McMillan & Singer 1993).

Repetitive elements are also seen in amino acid sequences. For example, immunoglobulins are composed of two copies of both heavy and light chains, which themselves share two basic sequence motifs. Multiple copies of this immunoglobulin domain are present in many cell surface receptors, cell adhesion molecules, and antibodies.

The study of repeats is important for a variety of reasons. The presence of repetitive elements makes sequence assembly difficult, since a unique ordering of the contigs may no longer exist. Polymorphic repeats are also used as physical and genetic markers. In addition, control and transcriptional elements are present in multiple copies in the genome, and a study of the repeats may aid in identifying them. Most importantly, repetitive elements may be used to infer the evolutionary history of the genome. The distribution of their prevalence and time of birth may pinpoint catastrophic events in the evolution of the genome.

The prevalence and diversity of repeats in the genome leads to a number of biological questions. How did these repetitive elements arise? What functions do they play? Are the repeat elements stable? To address these questions, tools are needed to recognize repetitive elements in anonymous sequence data, and to categorize and analyze the resulting families of sequence elements.

Recognizing repetitive sequence elements in genomic sequence is an example of a fundamental problem in machine learning, namely, pattern induction from an anonymous data stream. In the case of molecular sequence analysis, no prior knowledge of the length, location, number, or characteristics of the repeats is available. Therefore, general pattern induction methods need to be developed.

Milosavljevic and Jurka (1993) have used minimum length encodings to study significant repeats in a sequence, thus computing a short representation of a

sequence from an information-theoretic perspective. Fitch, Smith, and Breslow (1986) have examined the problem of detecting tandem repeats and analyzing their evolutionary history. A number of techniques also exist to identify exact repeats. A simple efficient algorithm utilizes suffix trees (Bieganski *et al.* 1994). Heuristic algorithms, like BLAST (Altschul *et al.* 1990) and FASTA (Pearson & Lipman 1988), also perform well in detecting repeats with few mutations. However, none of these fast searching techniques perform well for detecting short mutated repeats.

Blaisdell *et al.* (1993) have systematically studied significant repeats in an *E. coli* DNA sequence (length 1.6 million bases). Both small repeats with high copy number and larger repeats with smaller copy number are considered significant. They used an algorithm developed by Leung *et al.* (1991) for identifying repeats. This linear time algorithm relies on finding repeats with non-mutated blocks possibly connected by short mutated blocks. The sensitivity of this algorithm would appear to be similar to that of BLAST, and thus may not be adequate for discovering short, but significant mutated repeats.

We have adopted a sequence similarity scoring system based on a well-defined underlying statistical model (Altschul 1991; States, Gish, & Altschul 1991). Using this approach, we have been able to define statistically significant classes of repeated sequence elements. Furthermore, we have developed tools to analyze and visualize these classes, which aid in understanding the complicated relationships between the constituent elements of these classes. In particular, a hierarchical view of class relationships is developed to examine biological evolutionary relationships between elements.

Over 3.6 million contiguous base pairs of sequence from *C. elegans* chromosome III have recently been determined.¹ We have focused on this region for our analysis, since it includes one of the largest contiguous segment of DNA sequence available (Wilson and others 1994). Experimental studies with reassociation kinetics and electron microscopy have suggested that repetitive elements account for 17% of the *C. elegans* genome (Britten & Kohne 1968; Sulston & Brenner 1974; Wood 1988). Our computational analysis provides an estimate of approximately 12%. These repeat elements are diverse, with multiple apparent mechanisms of origin and evolution.

Methods

We use the term sequence *segment* to refer to a contiguous section of the DNA sequence. The term *repeat segment* refers to a sequence of DNA (of significant

¹This is almost a contiguous sequence, with 21 gaps ranging in size from 30,000 to 483,000 bases with the total gap size of approximately 2 million bases. Thus, these 3.66 million sequenced bases span a sequence of approximately 5.7 million bases.

length) that occurs more than once in the available sequenced genome. The *repeat segments* include segments both on the plus and minus strands of the DNA. The frequency of occurrence of a segment is referred to as its *copy number*. Thus, unique segments have copy number one, and repetitive elements have a copy number of at least two.

Significant score

A repeat segment is considered significant if it scores above a certain threshold. The threshold score is essentially $\log_2 N$ bits, where N is the search space (product of the lengths of query and the database sequence). For a self-similarity search of a sequence, N is approximately the square of the sequence length.

The statistical theory of maximal local alignments is well developed, and we judge significance according to Karlin and Altschul (1990).

A score S (in bits) is considered significant if:

$$S \geq \log_2 \frac{K}{p} + \log_2 N$$

K depends upon the substitution matrix, and was upper bounded by 0.4 for our choice of substitution matrices. p is the probability that the score S occurs by chance alone. We consider an alignment significant when p is 0.05. N is $L^2 + L(L - 1)/2$, where L is the total length of the examined sequence (3,655,029) for the *C. elegans*. A simple explanation for the value of N is that we are searching the lower triangle of the dynamic programming matrix for the plus strand against the plus strand, and the entire matrix for the plus strand against the minus strand, and the total number of starting points for possible alignments in these matrices is $L^2 + L(L - 1)/2$. Thus,

$$\begin{aligned} S &\geq \log_2 \frac{0.4}{0.05} + \log_2 (1.5L^2 - 0.5L) \\ &\geq 3 + \log_2 (2.0 \times 10^{13}) \\ &\geq 47.2 \text{ bits} \end{aligned}$$

In addition, since we use the best alignment score over all the PAM matrices (corresponding to an increase in search space), a correction of $\frac{1}{2 \times 0.693} \ln \ln N$ needs to be applied (Altschul 1993). This amounts to 2.5 bits, raising the cutoff to 49.7 bits. Therefore, we used 49.7 bits as the minimum score for a repeat to be considered significant.

However, for repeats with copy number greater than two, a smaller score may be considered significant (Altschul & Lipman 1990). If two copies of a repeat with a score of 50 bits are unlikely, then three copies of a repeat with a score of 36 bits are also unlikely. In general, for copy number equal to three, $N = 7L^3/6$, and a total score of 71.4 bits for the best two alignments of the three is significant².

²There are three search matrices to consider: all three

PAM distance

The uniform mutation model assumes that mutations occur with equal frequency over time and base position. The evolutionary distance between two sequence segments can be estimated using this model (States, Gish, & Altschul 1991). This mutation model provides a series of scoring matrices, each corresponding to an evolutionary distance. The evolutionary distance between two sequence segments is inferred from the number of the scoring matrix that maximized the information content (score) of the alignment. The uniform mutation model is an approximation, and there are some regions of DNA sequence where it does not hold; however, due to its simplicity it is still widely used.

PAM is an abbreviation for point accepted mutation. The PAM 1 scoring matrix maximizes the score for sequences, where each base has suffered an average of 0.01 mutations. The PAM number can also be regarded as a time unit. Thus, PAM 1 is the time taken for a sequence to achieve one point accepted mutation.

The various PAM matrices are built using a Markov mutation model. The diagonal probabilities for the one point accepted mutation probability matrix (M_1) are 0.99, and a biased mutation model provides the probability of transitions ($A \leftrightarrow G$ and $C \leftrightarrow T = 0.006$), and of transversions ($A \leftrightarrow C$, $A \leftrightarrow T$, $C \leftrightarrow G$, and $G \leftrightarrow T = 0.002$). The probability matrix corresponding to n PAM's is $M_n = (M_1)^n$. The element in the i^{th} row and j^{th} column of this matrix is M_{nij} . These probability matrices are converted to symmetrical log odds score matrices with scores S_{ij} 's. The score for aligning base i with base j at n PAM's is S_{nij} , and

$$S_{nij} + S_{nji} = \log_2 \frac{p_i M_{nij}}{p_i p_j} + \log_2 \frac{p_j M_{nji}}{p_j p_i}$$

p_i is the probability of occurrence of base i . For the *C. elegans* sequence ($p_A = p_T = 0.32$ and $p_C = p_G = 0.18$). Since the scores should be symmetrical,

$$S_{nij} = S_{nji} = \frac{1}{2} \log_2 \frac{M_{nij} M_{nji}}{p_i p_j}$$

Thus, we obtain a scoring matrix (with the scores in bits) for each PAM distance. Given a local alignment, one of the PAM matrices provides the best score for the alignment.³ The number of this *optimal* matrix is inferred to be the evolutionary distance between the two segments (involved in the local alignment).

segments on the positive strand ($N = L^3/6$), one segment on the positive strand and two on the negative strand ($N = L^3/2$), and two segments on the positive strand and one on the negative strand ($N = L^3/2$). In summation, the total search space is $N = 7L^3/6$.

³We use PAM matrices from 1 to 125. Though PAM matrices with distance larger than 125 can be used, they provide little information, and our tests failed to find any significant alignments at greater than PAM 125.

Local Alignment Algorithm

The problem of locating repeated segments is easily reduced to the problem of searching a sequence database for similarity. A number of heuristic tools exist for this purpose, including BLAST (Altschul *et al.* 1990), FASTA (Pearson & Lipman 1988) and FLASH (Califano & Rigoutsos 1993). Altschul *et al.* (1994) provide an excellent review of the issues involved in searching sequence databases.

BLAST We attempted to identify all local alignments by using BLASTN version 1.3.12 with option *-overlap* and word size $W = 8$. Since BLASTN is not guaranteed to work correctly with $W=8$, we also used $W = 12$, and merged the sets of alignments obtained. All the possible alignments indicated by BLAST were re-scored using the PAM matrices to establish a PAM distance for each alignment.

BLASTN is optimized to find nearly identical sequence segments rapidly. Our results indicated that it preferentially found alignments which had low PAM numbers. This was due both to the default scoring matrix it utilizes (+5 for a match, and -4 for a mismatch), and its requirement that at least W bases match exactly (where W is the word size). However, repeat segments with a high PAM distance may have a significant score without having an exact match of W bases.

FLASH is efficient for repeated searches on the same database; however, we required only a single search (Califano & Rigoutsos 1993). The code for FLASH is not available, thus it is difficult to use FLASH on the *C. elegans* sequence. We are currently evaluating the sensitivity and efficiency of FASTA for our repeat search.

Exhaustive search The limitations of BLASTN necessitated an exhaustive search for identifying all the local alignments. Since we were interested in only ungapped alignments, this exhaustive search is a special case of the dynamic programming algorithm (Waterman 1989). It can be easily implemented using $O(L)$ space. The problem is easily partitioned into sets of diagonals, which can be handled in parallel on different workstations. The sequence (L) being 3.6 Mb long, the dynamic programming was a major undertaking of $O(L^2) = (3.6 \times 10^6)^2 = 1.3 \times 10^{13}$ operations, which on a combination of 6 workstations (including SGI Indigo's, DEC Alpha 3000/500's, and SPARCstation 10's) took 4 days to compute.

We identified all the local alignments by scanning along each diagonal for the number of matches minus mismatches exceeding 10 (≈ 20 bits). We believe this number (10) is small enough to detect all the significant alignments. The set of local alignments is post-processed to weed out the alignments with insignificant scores. The best PAM distance and end points for each alignment are determined, and utilized as input for building the repeat graph.

Encoding repeat alignments as graphs

A graph $G = (V, E)$ is defined to capture the various relationships between the repeat segments. The set of vertices V is the set of repeat segments. The edges E represent alignments. The edges are weighted, and the edge weights correspond to the evolutionary distance between the sequence segments. There is no edge between unrelated segments.

Distinct families of repeats correspond to the connected components of this graph. In fact, the connected components that correspond to well-conserved repeat families should have high connectivity approaching that of cliques. The connected components can be identified algorithmically in $O(|E|)$ time. However, the identification of distinct vertices is somewhat arbitrary. The local alignment algorithm produces a pairwise list of sequence segments (along with a score and distance for each such pair). If we consider all such sequence segments, we get disjoint, overlapping, and totally included segments. We need a method to determine which overlapping sequence segments can be represented by the same vertex. If we merge segments into the same vertex generously, we may end up with one large connected component, and the various sequence segments involved will bear little similarity. We represent segments that have at least 90% overlap by a single vertex⁴ (i.e. $|A \cap B|/|A \cup B| > 0.9$). Thus, most vertices in a connected component will be of similar length.

Each connected component (C) corresponding to a repeat family is reduced to a minimum spanning tree (T). The minimum spanning tree presents a succinct encoding of the evolutionary relationship between segments (States, Harris, & Hunter 1993).

Consider a single connected component. We use Kruskal’s algorithm for building the minimum spanning tree (Corman, Leiserson, & Rivest 1990). Kruskal’s algorithm sorts the edges by non-decreasing weight. It considers each edge in order and if the vertices on this edge are not already connected, it includes this edge in the spanning tree. This has the advantage that the set of trees obtained after the edges with weight less than w have been considered correspond to all the evolutionary relationships in the family within w PAM’s. Each such tree (T_w) is reduced to a single vertex, which represents a common ancestral sequence for the vertices (sequence segments) in T_w .⁵

In other words, we build this tree PAM level by PAM level. Consider figure 1. The entire sequence is laid out along the x-axis (the i^{th} base is at $x = i$). The PAM distance is along the y-axis. Each set of vertices (sequence segments) that are connected by edges with

⁴This choice is heuristically motivated, and different overlap criterion provide somewhat different family structures and total estimates of information duplication.

⁵In fact, these spanning trees are a form of Steiner trees, as we are introducing additional vertices (corresponding to ancestral sequences) into the graph.

weight PAM 1 in the spanning tree is joined by edges to a common point (an additional vertex is introduced) at PAM 1, indicating that they evolved from a common unknown ancestor within one PAM.⁶

Efficient implementations (using Fibonacci heaps) of the minimum spanning tree have a complexity of $O(|E| + |V| \log |V|)$, and even simple implementations provide $O(|E| \log |E|)$ running time (Corman, Leiserson, & Rivest 1990). Moreover, since our edge weights are integral and bounded (by the minimum and maximum PAM distance), we can use Binsort to sort the edges to achieve a simple $O(|E| + |V| \log |V|)$ running time algorithm. The bottleneck, however, is not the minimum spanning tree implementation but the repeat discovering step.

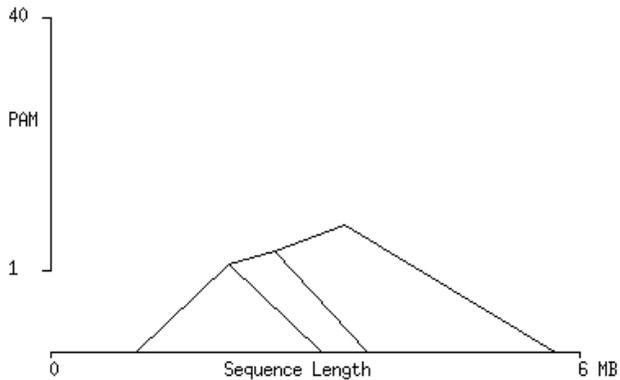
Results

A salient point of our analysis was the number of repetitive segments and the diversity of families discovered⁷. An assessment of the duplication reveals that the sequence repeats account for approximately 12.3% of the region of the *C. elegans* genome. Figure 2 exhibits the regions of sequence duplication along the analyzed segment of chromosome III. The sequence is laid on the x-axis, and the number of the bases in each block of 10 kilobases that participate in a significant repeat are plotted on the y-axis. Most of the extensive regions of the chromosome that appear unique are unsequenced regions for which no data was available. The regions of high duplication activity correspond to a combination of large local repeats (the largest was almost 8 kb), tandem repeats and local inverted repeats. However, the vast majority of the repeats were scattered randomly over the chromosome. In summary, about 450 kb of the 3,655 kb are involved in a duplication event, corresponding to 12.3% of the analyzed sequence.

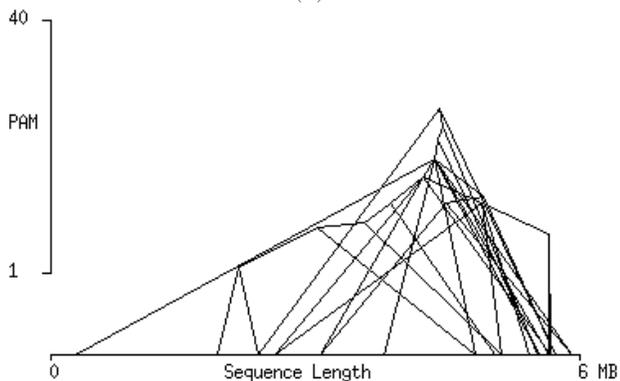
Figure 3 plots the distribution of the sizes of the various repeat families. About half the families had a copy number of two, and a fourth of the families had a copy number of three. This suggests that many repeat segments were missed because only a single copy (or two copies of a repeat that may only be three way significant) were present in the analyzed region of sequence. The region of chromosome III analyzed represents about 3.6% of the total *C. elegans* genome. In addition, one of the gaps in the sequence is known to be composed almost entirely of repetitive elements. As more sequence data becomes available, the significance level for a repeat will rise. Thus, some of the current

⁶This visualization imposes an ultrametric on the evolutionary distances. If a segment A is related to two other segments B and C by distance of PAM d , then B and C are assumed to be related by a distance of at most PAM d . Thus, each segment has only a single edge emanating from it, forcing the evolutionary structure to be a tree rather than a graph.

⁷Internal structure of tandem repeats was excluded from the analysis.



(a)



(b)

Figure 1: Spanning trees representing evolution. x range is the entire contiguous sequence about 6 million bases. (a) A small repeat family with 4 members. Two members are separated by PAM 1, and these are connected to a newly introduced vertex at PAM 1. The x location of the new node is chosen to be the weighted midpoint of the sequence segments involved. A third member is connected to the ancestor of the first two members at PAM 3, and the last member is connected to the ancestor of the previous three family members at PAM 7. (b) A larger repeat family with members all over the chromosome.

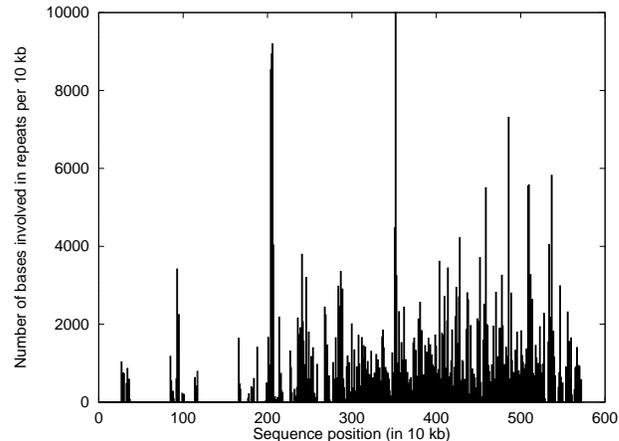


Figure 2: Extent of sequence duplication. The sequence is laid out in 10 kb blocks along the x-axis. On the y-axis is plotted the number of bases of each 10 kb block, which are involved in a duplication event. For reference, the 2.2 Mb of contiguous data reported by Wilson et al. (1994) is the sequence from 2.9 Mb to 5.1 Mb in our numbering scheme.

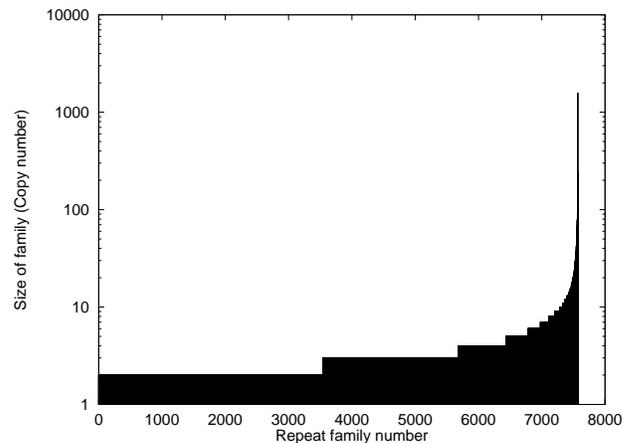


Figure 3: The number of repeat segments in each family. The repeat families were sorted according to increasing size before plotting.

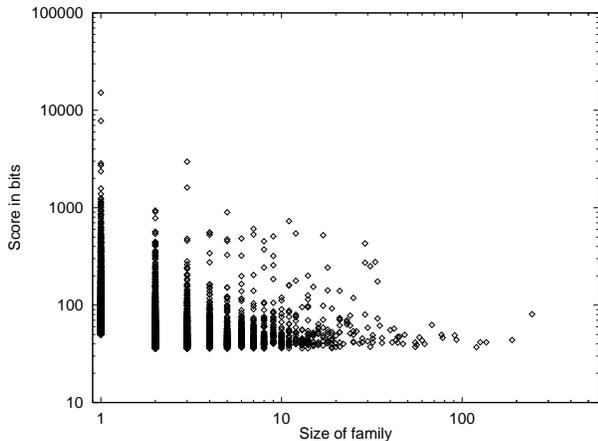


Figure 4: A scatter plot of the average score in a repeat family as a function of the size of the family.

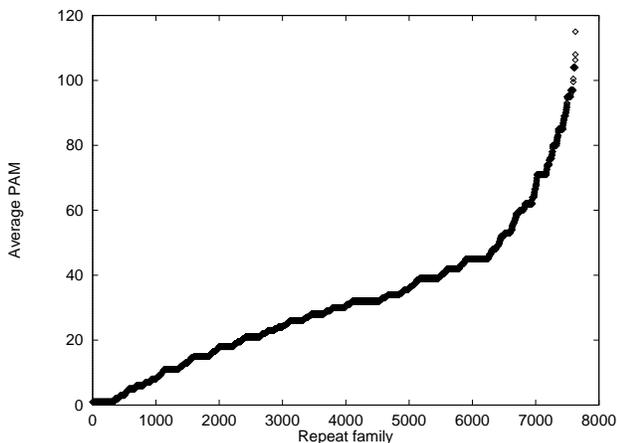


Figure 5: Evolutionary divergence (average PAM) in each repeat family. The repeat families were renumbered in order of increasing average PAM distance.

borderline repeats will be no longer significant, but this effect will be overshadowed by the discovery of new repeats, and we believe that the estimate for the repeat content of the *C. elegans* genome will increase from the current figure of 12%.

A scatter plot of the average score in a family versus the size of the family in figure 4 reveals that small families have large variation in the size of the segments involved in the duplication. The significance levels of the scores are also evident from the minimum score for a family. Families with size three or more require a lower score to be considered significant due to a three-way alignment. Families with large copy numbers tend to be composed of small repeats.

The date of a duplication event can be estimated from the number of substitution events that have accumulated between copies. Figure 5 shows the distribution of evolutionary divergence as a function of the

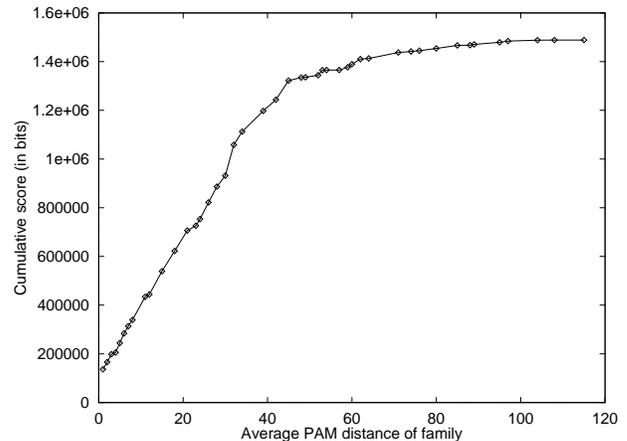


Figure 6: The cumulative amount of information present in all repeats for a given PAM.

repeat family. A remarkably linear distribution was observed up to approximately 40 point mutations per hundred nucleotides (PAM). These results suggest that sequence duplication events have been occurring at a constant rate in recent evolution. The decline in the number of repeat families discovered beyond 40 PAMs probably represents the limitations of our search technique rather than an intrinsic biological process. The search technique is limited by both the small amount of information available per base in the higher PAM scoring matrices, and our inability to detect repeat alignment with high PAMs, since we ignore insertion and deletion mutations. At 40 PAMs, an alignment on the average provides 0.67 bits per aligned nucleotide. To achieve statistical significance of 50 bits, a duplication must be at least 75 nucleotides long, and even though about 30 (40% of 75) substitution mutations must have taken place, yet no insertion or deletion mutation must have occurred.

The constant rate of production of repeats is reinforced by figure 6, which shows the cumulative information content in the repeat families as a function of the average PAM distance of the family. As expected, the information content saturates rapidly with increasing PAM distance. In fact, over 90% of the repeat information is present within an evolutionary distance of PAM 50. An evolutionary distance of PAM 1 corresponds to about 100,000 to 1 million years⁸ Thus, duplication events within the past 5–50 million years account for at least 12% of the *C. elegans* genome.

Figure 7 shows a repeat family with 22 members. This is a local repeat family with each sequence segment about 85 bases. The entire family is contained

⁸This assumes a spontaneous mutation rate of 10^{-9} to 10^{-10} point substitutions per base per generation. Assuming 100 generations per year, this corresponds to 10^{-7} to 10^{-8} mutations/base/year or 10^5 to 10^6 years for 1 mutation amongst 100 bases. This is only a ballpark number.

within a section of 2,500 bases. Such repeat families can be accounted by a local diffusion phenomena. Figure 1 exhibits two other representative repeat families. These have members spread over the entire 6 Mb region, and these cannot be accounted by a model of local rearrangement.

Discussion

The distribution of family size as a function of family number exhibits exponential characteristics. 75% of the total families (7629) had only a single or a double duplication event (copy number two or three) in the analyzed region of chromosome III (figure 3). If we extrapolate this distribution to the entire genome, it is likely that we will discover more members of these families and other new families. The 3.66 Mb that we have examined is only a third of the length of chromosome III, and only a thirtieth of the *C. elegans* genome. We expect to discover greater than 12% repeats, when the entire genomic sequence becomes available and is examined.

The number of true families of repeats is probably smaller than the number we discovered (7629). Accounting for insertion and deletion mutations will reduce the number of families. Manual examination and annotation of the families will also correct errors due to the heuristic utilized to form families.

A linear correlation was observed between the evolutionary divergence of the family members and the cumulative number of repeat families (figure 5). The sensitivity of our similarity search breaks down at approximately 40 PAMs. For non-coding genomic sequence, 40 PAMs corresponds to a relatively short time period (about 4–40 million years). For example, *C. briggsae* and *C. elegans* are thought to have diverged approximately 30 million years ago, but share no recognizable sequence similarity in homologous non-coding regions of the genome (Emmons, Klass, & Hirsh 1979). Therefore, most of the non-coding duplication events that we have observed are likely to have occurred in the last 30 to 40 million years. Events leading to the introduction of repeat families with size 30 or more per genome must occur every few thousand years, and events leading to families with small sizes may be quite frequent.

Duplication events occurring in the last 30 to 40 million years account for about 12% of the information in the analyzed segment of chromosome III. Extrapolating the amount of repetitive sequence in the genome back in evolutionary time, duplication events occurring after the radiation of the major animal phyla about 600 million years ago could account for much of the information present in the *C. elegans* genome. Obviously, protein coding regions are being ignored in this analysis. Nevertheless, our results suggest that the majority of the non-coding sequence in the *C. elegans* genome is derived from duplication events that have occurred since the higher animals diverged from each other. However, most of the evolution events that

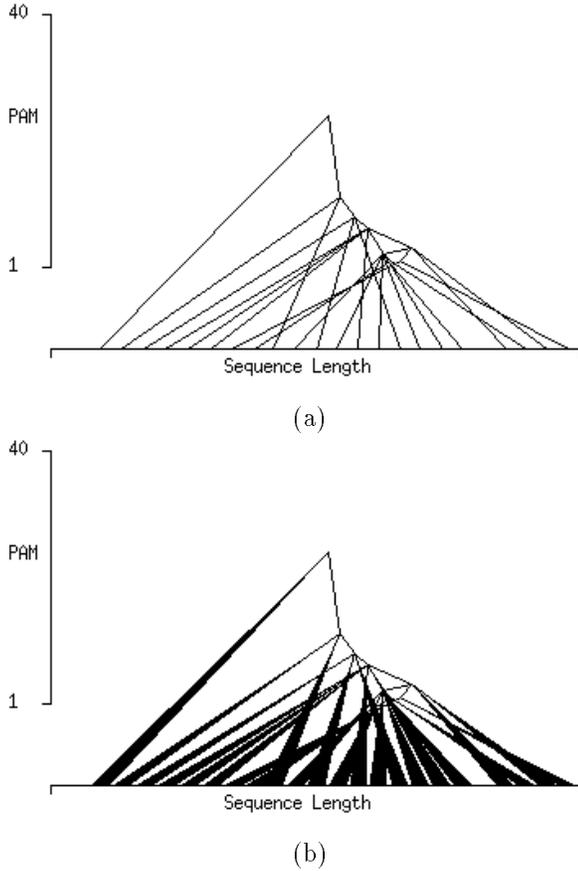


Figure 7: Local evolutionary tree. The range of x is 2500 bases, and its offset is 4.042Mb; thus, this entire repeat family is present between 4,042,000 and 4,044,500 bases (a) Only edges from the midpoints of the repeat segments are drawn. (b) The extent of the sequence involved in the duplication is shown. Notice a majority of the 2500 base segment is covered by 22 repeat segments, each about 85 bases long.

have occurred more than 40 million years ago cannot be identified.

The fact that genome sizes of phylogenetically related species are similar suggests that the overall genome sizes have been relatively stable, although the possibility of a uniform growth in genome size among all contemporary members of the phyla cannot be excluded. To maintain a stable genome size in the face of a constant rate of introduction of new sequence, sequence loss must be occurring at a comparable rate.

Our analysis techniques differ from those employed by Blaisdell et al. (1993). We employ a more rigorous, but computationally expensive search technique. We use standardized significance tests employing a bit-scoring criterion rather than a number of bases matching-mismatching criterion. The focus of our work is on estimating the total repeat content and evaluating the evolutionary implications; while Blaisdell et al. annotated the repeat families they obtained.

A limitation of the current analysis is our inability to consider repeats that have undergone insertion or deletion mutations (indels). Repeats with indels cause problems in three of the constituents of our toolkit: judging significance, computing evolutionary distance, and locating the repeats. The statistical theory provided by Karlin and Altschul (1990) does not extend to gapped alignments. It is also more difficult to judge evolutionary distances between repeats that have suffered indels. There are also reasons of computational efficiency. Unfortunately, our tests with computationally efficient search tools such as BLASTN have showed that they missed at least some distantly related sequence duplications. We have therefore used a full sequence comparison with a tightly coded inner loop, yet, the analysis of chromosome III required several days of compute time on a farm of workstations operating in parallel. Using a linear space implementation of the dynamic programming algorithm would allow for the presence of insertions and deletions, but the inner loop is more complex, and the overall calculation would be more time consuming (Waterman & Eggert 1987; Myers & Miller 1988). We are also examining parallel computing solutions for the search problem. We expect that the use of more sophisticated search techniques will improve the sensitivity of our similarity search, but we do not expect it to alter our results qualitatively.

Acknowledgments We would like to thank Hugh Chou, D'vorah Graeser, Toni Kazic, Lauren Treacy and the members of the *C. elegans* sequencing center at Washington University for their suggestions and comments.

References

Altschul, S., and Lipman, D. 1990. Protein database searches for multiple alignments. *Proc. Natl. Acad. Sci. USA* 87:5509-13.

Altschul, S.; Gish, W.; Miller, W.; Myers, E.; and Lipman, D. 1990. A basic local alignment search tool. *J. Mol. Biol.* 215:403-410.

Altschul, S.; Boguski, M.; Gish, W.; and Wootton, J. 1994. Issues in searching molecular sequence databases. *Nature Genetics* 6:119-129.

Altschul, S. 1991. Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.* 219:555-565.

Altschul, S. 1993. A protein alignment scoring system sensitive at all evolutionary distances. *J. Mol. Evol.* 36:290-300.

Bieganski, P.; Riedl, J.; Carlis, J.; and Retzel, E. 1994. Generalized suffix trees for biological sequence data: Applications and implementation. In *Proceedings 27th Annual Hawaii International Conference on System Sciences*, 35-44. IEEE.

Blaisdell, B.; Rudd, K.; Matin, A.; and Karlin, S. 1993. Significant dispersed recurrent DNA sequences in the *Escherichia coli* genome: Several new groups. *J. Mol. Biol.* 229:833-848.

Britten, R., and Kohne, D. 1968. Repeated sequences in DNA. *Science* 161(3841):529-540.

Britten, R.; Baron, W.; Stout, D.; and Davidson, E. 1988. Sources and evolution of human *Alu* repeated sequence. *Proc. Natl. Acad. Sci. USA* 85:4770-4774.

Califano, A., and Rigoutsos, I. 1993. FLASH: a fast look-up algorithm for string homology. In Hunter, L.; Searls, D.; and Shavlik, J., eds., *Proceedings First International Conference on Intelligent Systems for Molecular Biology*, 56-64. Menlo Park: AAAI Press.

Corman, T.; Leiserson, C.; and Rivest, R. 1990. *Introduction to Algorithms*. The MIT Press.

Deininger, P., and Schmid, C. 1979. A study of the evolution of repeated DNA sequences in primates and the existence of a new class of repetitive sequences in primates. *J. Mol. Biol.* 127(4):437-60.

Emmons, S.; Klass, M.; and Hirsh, D. 1979. Analysis of the constancy of DNA sequences during development and evolution of the nematode *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA* 76(3):1333-1337.

Fitch, W.; Smith, T.; and Breslow, J. 1986. Detecting internally repeated sequences and inferring the history of duplication. In *Methods in Enzymology*, volume 128. Academic Press. 773-788.

Jurka, J.; Walichiewicz, J.; and Milosavljevic, A. 1992. Prototypic sequences for human repetitive DNA. *J. Mol. Evol.* 35:286-291.

Karlin, S., and Altschul, S. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA* 87:2264-2268.

Leung, M.-Y.; Blaisdell, B.; Burge, C.; and Karlin, S. 1991. An efficient algorithm for identifying matches

- with errors in multiple long molecular sequences. *J. Mol. Biol.* 221:1367–1378.
- McMillan, J., and Singer, M. 1993. Translation of the human LINE-1 element, L1Hs. *Proc. Natl. Acad. Sci. USA* 90(24):11533–11537.
- Milosavljevic, A., and Jurka, J. 1993. Discovering simple DNA sequences by the algorithmic significance method. *Comput. Appl. Biosci.* 9:407–411.
- Myers, E., and Miller, W. 1988. Optimal alignments in linear space. *Comput. Appl. Biosci.* 4(1):11–17.
- Pearson, W., and Lipman, D. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* 85:2444–2448.
- States, D.; Gish, W.; and Altschul, S. 1991. Improved sensitivity of nucleic acid database searches using application-specific scoring matrices. *Methods: A Companion to Methods in Enzymology* 3(1):66–70.
- States, D.; Harris, N.; and Hunter, L. 1993. Computationally efficient cluster representation in molecular megaclassification. In Hunter, L.; Searls, D.; and Shavlik, J., eds., *Proceedings First International Conference on Intelligent Systems for Molecular Biology*, 387–394. Menlo Park: AAAI Press.
- Sulston, J., and Brenner, S. 1974. The DNA of *Caenorhabditis elegans*. *Genetics* 77:95–104.
- Waterman, M., and Eggert, M. 1987. A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. *J. Mol. Biol.* 197(4):723–728.
- Waterman, M. 1989. *Mathematical methods for DNA sequences*. Boca Raton: CRC Press.
- Wilson R. and others. 1994. 2.2 Mb of contiguous nucleotide sequence from chromosome III of *C.elegans*. *Nature* 368:32–38.
- Wood, W. 1988. *The Nematode Caenorhabditis elegans*. New York: Cold Spring Harbor Laboratory Press.