CABIOS

Vol 11 no 6 1995
Pages 615–619

Miropeats: graphical DNA sequence comparisons

J.D.Parsons

Abstract

Miropeats displays DNA sequence similarity information graphically. The program discovers regions of similarity amongst any set of DNA sequences and then draws a graphic that summarizes the length, location and relative orientations of any repeated sequences. Sequence similarity searching is a very general tool that forms the basis of many different biological sequence analyses but it is limited by the verbosity of traditional alignment presentation styles. Miropeats enhances the utility of conventional DNA sequence comparisons when looking at long lengths of sequence similarity by summarizing large-scale sequence similarities on a single page of PostScript graphics. Miropeats has been applied estensively to help understand shotgun assembly projects, to check cosmid overlaps and to perform inter-genomic comparisons.

Introduction

DNA sequencing projects around the world are producing sequence information at an accelerating rate and discovering that existing sequence analysis and comparison tools are often inadequate for the volume of data being handled. Miropeats is a novel program, designed from the outset to provide the user with graphical summaries of extensive and sometimes complex sets of DNA sequence similarity information. The descriptive abilities of Miropeats open research opportunities that would not be possible, or would be difficult to do otherwise. Examples include comparing the repeat structures of entire chromosomes, visualizing overlapping sequence fragments in a contig assembly project and comparing the consensus sequences produced by different contig assembly programs. Miropeats was originally written to help contig assembly projects at the Genome Sequencing Center in St Louis, USA, where it was found to be useful for this and may other different roles. The intrinsic inscrutability of a string of 40 000 characters picked from an alphabet of only four letters (a typical cosmid assembly project) is made worse because the shotgun sequencing strategy starts with the original contiguous 40 Kb DNA sequence split into 700 overlapping, but imperfectly copied, pieces that must be joined back together. Miropeats helps shotgun assembly, not by performing the assembly itself, but by helping the researcher gain an overall understanding of the task left remaining after an initial round of fragment assembly using whatever assembly program the user is familiar with. Miropeats can do this because it draws a simple PostScript graphic that shows potential joins, cosmid overlaps, and also distinguishes tandem repeats, inverted repeats, oligo repeats and palindromes from each other. (PostScript is a trademark of Adobe Systems Inc.)

Genome Sequencing Center, Washington University School of Medicine 4444 Forest Park Avenue, St Louis, MO 63108 USA.
Email: jparsons@watson.wustl.edu

Algorithm and implementation

Miropeats itself, is a single UNIX C-shell script; all the DNA comparisons are done by calls to another program called ICAass which is written in ANSI-C. ICAass is one of the latest ICAtools (Parsons *et al.*, 1992) and is fully described in an accompanying paper. ICAass compiles an index of all those sequences to be examined and then compares each of them against the others and itself in both orientations looking for repeated segments. Any matching ungapped segments with a match score higher than a defined threshold are reported as a list of easily parsed alignments. The Miropeats script just parses out the position and quality of the matching DNA segments and then coverts this information into PostScript graphics. ICAass was chosen as the DNA comparison engine because its output is extremely easy to parse, it has a good compromise between sensitivity and speed, and because it handles many DNA sequence formats. Graphical studies of DNA sequences have been attempted before, for example by Searls (1993), but this work was limited to quick previews of short DNA sequences. Searls' program is constrained primarily because it uses PostScript as a general programming language running on the limited resources of the printer itself to do both its sequence analyses and page layout calculations.

The objects in the graphical display are: DNA sequences represented as thin horizontal lines starting at the left margin and terminating on the right hand side in a solid triangle; repeat regions indicated by thick black lines overlaid on the sequence; and thin curving lines that link the termini of the equivalent ends of the two copies of a

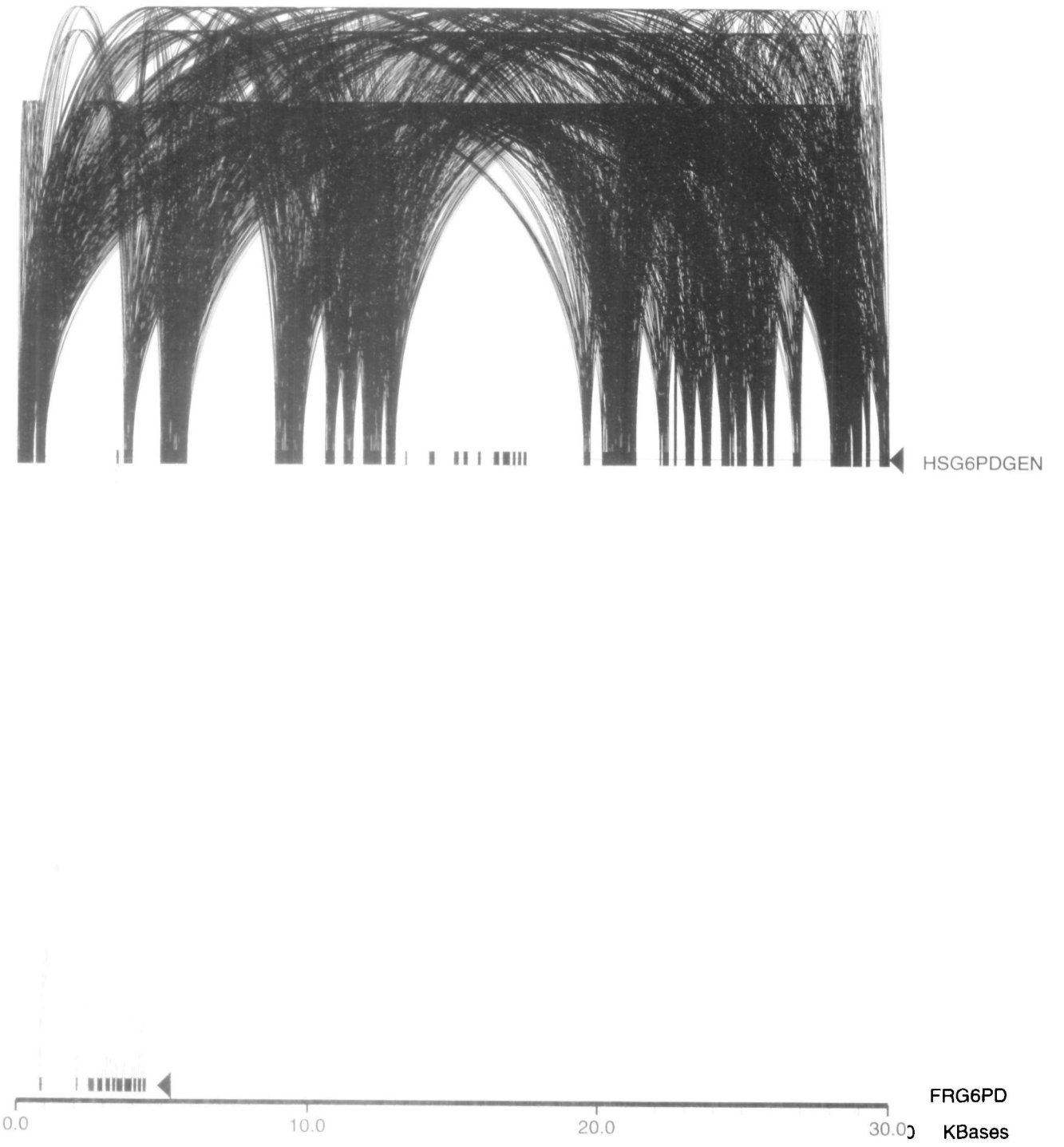

Fig. 1. A comparison between two instances of a homologous gene, one is from *Fugu rubripes* (GenBank: FRG6PD) and the other is Human (Genbank: HSG6PDGEN), as performed by Miropeats. Only the first 30 000 bases of the HSG6PDGEN entry that correspond to the glucose-6-phosphate dehydrogenase gene were used The numerous lines drawn above the human sequence are mostly links joining pairs of Alu repeats. These repeats are obviously absent from the Fugu fish homologue The 11 pairs of lines that joint the two G6PD sequences correspond to the 11 of the 12 exons that are found in both homologues of this gene. Miropeats was run with the default threshold of 30.

repeat. Most of the graphical work is done using high level drawing commands built into the PostScript language. Those linking lines that join the ends of repeats on the same piece of sequence need special processing in order to raise the middle of the linking lines above their origins to make the links visible. By considering the two ends of a repeat separately, it is possible to use different apogees for different types of

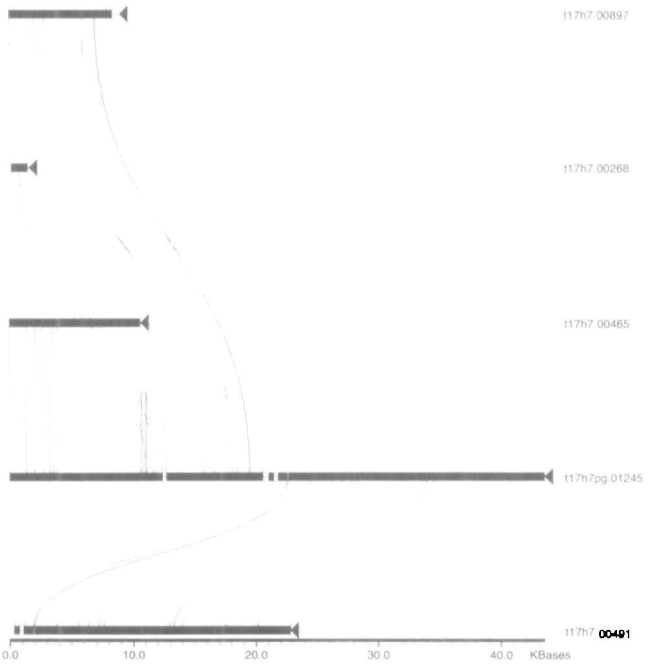

Fig. 2. A comparison between the contig sets produced by two different assembly programs when given the same set of shotgun reads. It is obvious that the contigs do not form simple pairings so there is disagreement between the two programs in the way that the reads should be assembled together. Phil Green's program Phrap (unpublished) produced a single contig (2nd from bottom) whereas Bap (Dear,S. and Staden,R., 1991) produced four contigs, including one (t17h7 bap.00897) which was discovered to be incorrectly assembled internally.

repeat on the different ends of one repeat to enable the different repeat types to be distinguished in a relatively intuitive way (see Figure 3).

Miropeats has options to look at all repeated DNA sequence segments (default) or only those repeated sequences where either both copies are on a single sequence, or where both copies are on different sequences. The program also has an adjustable threshold which allows the user to choose what amount of DNA sequence similarity should be considered significant and therefore worth displaying. This facility allows Miropeats to be used for analysing different features in sequences varying from less than Kilobase to more than a Megabase. If the picture is too complex then the threshold can be raised or, if the picture is not displaying some repeats of interest, then the threshold can be lowered. The score of any matching segments is defined as the number of matching bases minus the number of non-matching bases. No account is taken of any possible matches between the sequences if insertions or deletions were included.

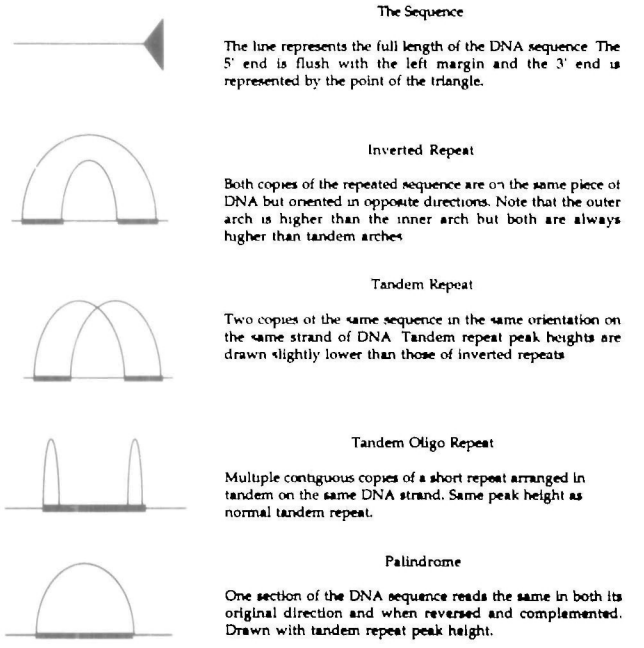

Fig. 3. A key to understanding Miropeats graphics where both repeat copies are on a single sequence

Sequence format requirements

Miropeats can use DNA sequences stored in any of six formats: EMBL, GenBank, FASTA, Staden Experiment, Staden Seqfile, or plain format. The first five listed are complex formats and it is possible to have any number of sequences of the same format in any one file. All the sequence data in plain (unformatted) files is assumed to come from a single sequence. As currently configured, ICAass will work with any number of sequences up to 4 Megabases in length. Miropeats was written and tested on Solaris 2.3 so the script may need slight alterations to run on different UNIX versions. All software mentioned in this paper is available via anonymous ftp from genome.wustl.edu in the file /pub/gsc1/parsons/miropeats.tar.Z.

Results

Intergenome comparison

Figure 1 shows a comparison between two instances of a homologous gene, one is from *Fugu rubripes* and the other is human, as performed by Miropeats. It is obvious how much more compact the *Fugu* gene is compared to the human homologue.

Sequence assembly comparisons

It is often useful to be able to compare the contigs produced by different assembly programs when given

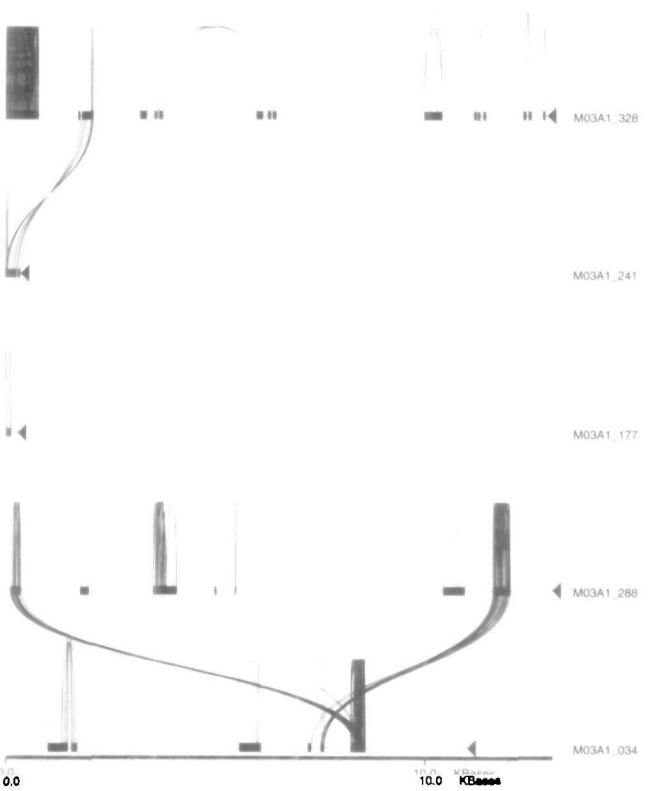

Fig. 4. An example of what a relatively complex *C. elegans* sequence assembly project can look like after an initial round of shotgun assembly. Many features are represented including examples of each of the different types of DNA repeats distinguishable using Miropeats: tandem, inverted, tandem oligo, and palindrome. The graphic also shows that contig number '00241' is probably just a subsequence of '00328' that has not yet been integrated into a single contig See Figure 3 for a graphic key

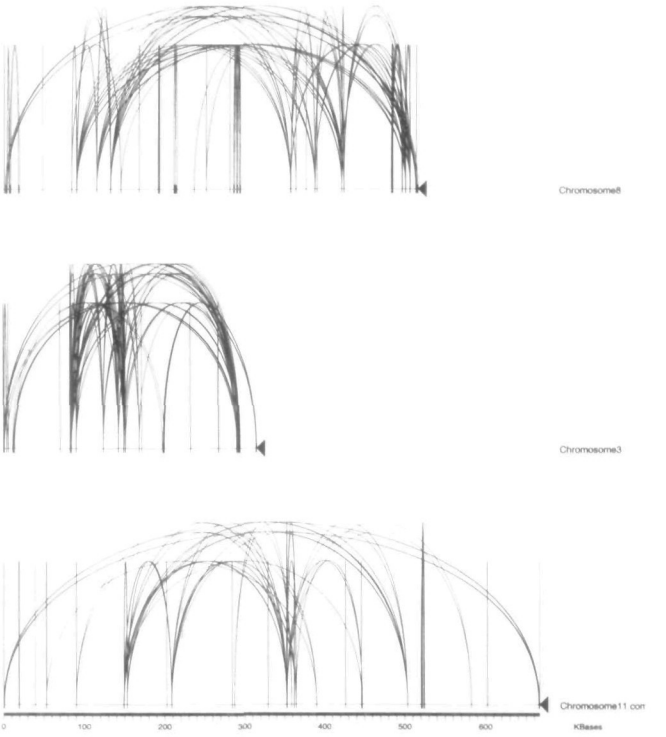

Fig. 5. Graphic showing a Miropeats analysis of three complete yeast chromosomes showing only internal repeats. Threshold was set to 40. All chromosomes are oriented with their short arms on the left side of the page. The telomeres are obvious as inverted repeats in all the chromosomes, the mating locus is the main feature in the centre of Chromosome 3 (at ~ 200 Kb) and two Hexose transport gene duplications are shown next to each other in Chromosome 8 (at ~ 300 Kb).

the same set of shotgun sequencing reads. Differences in the contig sets suggest that at least one of the assemblies is incorrect and will need attention. Figure 2 shows an example of this type of comparison performed by Miropeats and reveals one obvious area of disagreement.

Sequence assembly project views

An example of what a relatively complex *C. elegans* sequence assembly project can look like when some initial 'finishing' work has been done is presented in Figure 4. Many features are represented including examples of each of the different types of DNA repeats distinguishable using Miropeats. To help understand how Miropeats represents those repeats hat have both copies on a single piece of DNA see the key in Figure 3.

Chromosome analysis

Most DNA analysis programs could not read or sensibly display entire yeast chromosomes of over 600 Kb length but Miropeats can do so quickly and usefully. It takes ~ 4 min on a 40 MHz SPARCstation to examine yeast chromosome 8 (570 Kb) looking for all its internal repeats of at least 30 bases length. To see an example showing three separate chromosomes on a single diagram, see Figure 5. Each chromosome's internal repeat structure is made more obvious by opting not to draw links between repeats on different chromosomes.

Conclusions

Miropeats in its present form should be useful to any laboratory sequencing or analysing large DNA sequences. The program could still be improved: it is not very sophisticated about the positioning of sequence fragments on the page and this can lead to interesting information behind hidden behind a jumble of crossing lines. Furthermore, when working with Human DNA sequences or any DNA containing mostly repeats, there is a chance that interesting matches will be lost amongst the many Alu's etc. that are not usually of much scientific consequence. It should be possible to use one of the existing alternatives for repeat sequence masking

(e.g. Claverie and States, 1993) to circumvent this problem but none have been tested yet.

References

Claverie,J.M. and States,D. (1993) Information enhancement methods for large scale sequence analysis. *Computers Chem.*, **17**, 191–201.

Dear,S. and Staden,R. (1991) A sequence assembly and editing program for efficient management of large projects. *Nucleic Acids Res.*, **19**, 3907–3911.

Parsons,J.D., Brenner,S. and Bishop,M.J. (1992) Clustering cDNA sequences. *Comput. Applic. Biosci.*, **8**, 461–466.

Searls,D.B. (1993) Doing sequence analysis with your printer. *Comput Applic. Biosci.*, **9**, 421–246.

Received on May 1, 1995; accepted on September 13, 1995