Final Report On Research Grant DE-FG0395ER62139.

Identification, Organization and Analysis of Mammalian Repetitive DNA Jerzy Jurka

The three major objectives of this research were: (a) identification, (b) organization, and (c) analysis of mammalian repetitive DNA, as stated in the title. Our approach was based on computer-assisted analysis of sequence data complemented by small-scale experimental verification. Our major accomplishments include development of public database of repeats (Repbase Update) and of a server for automatic annotation of repetitive DNA, known as "CENSOR server". Furthermore, we have made a major contribution towards understanding of the mechanism of retrotransposition of SINE and LINE elements which represent 25% of mammalian genomes.

We have identified and characterized several dozen new families of repetitive elements from mammals published both in printed journals (1,2) and electronically in Repbase Update (3). This grant was the only funding which supported the development of Repbase Update throughout most of 1996. Repbase Update became the major resource on genomic repetitive DNA and is being used in genomic analyses worldwide. Repbase Update is lasting contribution to genome projects and it owes much of its existence to the DOE funding.

We have developed a new version of CENSOR Server based on dedicated hardware (Fast Data Finder from Paracel Inc.), run in conjunction with a Sun Ultra 10 (4). Our server replaced Pythia operations previously available from Argonne National Laboratories (5,6). CENSOR server continues to operate after expiration of this grant. It was used to annotate repeats in over 200,000,000 bp of mammalian sequence data submitted from all over the world. CENSOR also served as cross-reference software for developing RepeatMasker (Smit, A.F.A., 1996). CENSOR is particularly useful for annotation of long chunks of DNA sequences (over 100 kb), which often cause RepeatMasker to crash.

Using the new CENSOR 2.0 software, we have generated the first Genbank-scale map of repetitive elements based on available genomic DNA sequences from primates (7). The original map based on GenBank 105.0 contains over 200,000 repetitive elements and was partially supported by this grant. The major families of repeats are L1 (13.71%) and Alu (10.5%). The third largest group are various viral elements (7.5%), followed by L2 (1.9%), simple repeats (1.6%), and MIRs (1.5%). The remaining 1.9% are DNA transposons, satellites and unclassified elements. The map includes the following information: (i) accession numbers; (ii-iii) beginning and ending position of the repeat; (iv) repeat name as listed in Repbase Update; (v-vi) positions of the consensus sequence aligned to the particular repeat; (vii) orientation of the repeat - d)irect or c)omplementary; (viii) relative similarity to consensus sequence and (ix) ratio of mismatches to transitions. Given adequate resources such maps can be revised and updated on a regular basis. Currently, over 300 megabases (Mb) of primate DNA sequences have been deposited in Genbank 112.0. Generation of repeat annotations continues after expiration of this grant and is likely to become a standard feature in genomic databases. The current map contains over 600,000 repeats and repeat fragments and is available over the internet.

Our major research accomplishment was identification and characterization of sequence targets for integration of non-LTR retroelements (retroposons). These targets appear to be common for SINE elements such as Alu, B1, B2, BC1, as well as for L1 (LINE1) and processed retropseudogenes (8-11). This, and identification of endonucleolytic cleavage of similar targets by L1-encoded reverse transcriptase (12) established a foundation for understanding of the

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, make any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

mechanism of retroposition in mammals (10,12). The proposed mechanism may be applicable to non-LTR retroelements not only in mammals and other animals but also in plants (13).

Currently, we are studying potential relationships between targets for retroposon integration and recombinational hot spots in mammalian chromosomes (14). We have demonstrated in vitro that targets for retroposon integration are hot spots for homologous recombination, if present in multiple tandem copies. This has been demonstrated in plasmids transfected to mammalian cell lines lacking p53 tumor suppressor protein, such as C33A cell line. Co-transfection of p53 gene to C33A inhibited the recombination. Genomic sequences targeted by retroposable elements can become hot spots for homologous recombination in p53-deficient cancer cells whose p53 genes are deleted or mutated, p53 proteins are inhibited by viral proteins (e.g. SV40 T antigen) or endogenous oncoproteins (e.g. MDM2). This phenomenon can be used for cancer- targeted gene therapy if the overall frequency of recombination can be improved by another factor of magnitude. It also has implications for understanding of the overall stability of the genome.

Personnel involved: Paul Klonowski (programmer), Vladimir Kapitonov (post-doctoral fellow), and Jerzy Jurka P.I. (all 3 funded approximately until the end of 1996). Since 1997 we had only a limited funding for a technician Jiong Ma, working on a pilot experimental program devoted to analysis of the integration targets, as described above. This small-scale experimental project continues, and we hope to complete it at the end of 1999 using non-DOE funding.

Funding history: This award has originally been granted to Linus Pauling Institute in April of 1994 as a 3-year extension of the DOE grant DE-FG03-91ER61152. After approximately one year of operation at Linus Pauling Institute, the balance has been transferred to Genetic Information Research Institute under a new number DE-FG0395ER62139. The funding continued through March 31, 1997. After that we have applied for no-cost extension and a small supplemental funding to support our experimental work for one year. The supplemental funding for one lab technician was granted in 1998 and expired on March 31st, 1999.

1. Jurka, J., Kapitonov, V.V., Klonowski, P., Walichiewicz, J., Smit, A.F.A. *Genetica* **98**, 235- 247 (1996).
2. Kapitonov, V.V., Jurka, J. *DNA Sequence* **8**, 277-288 (1998).
3. Repbase Update: <http://www.girinst.org/~server/replibase.html>.
4. Klonowski, P., Jurka, J.: CENSOR Server version 2.0: cursor@charon.girinst.org.
5. Jurka, J., Walichiewicz, J., Milosavljevic, A. *J. Mol. Evol.* **35**, 286-291 (1992).
6. Milosavljevic, A. *Repeat Analysis. ICRF Handbook of Genome Analysis*, Blackwell Scientific Publications, Chapter 25, pp. 617-628 (1998).
7. Klonowski, P., Jurka, J.: <http://www.girinst.org/public> (1998).
8. Jurka, J. Site-directed Recombination. U.S. Patent 5,695,977 (filed in 1995; issued in 1997).
9. Jurka, J. *J. Mol. Evol.* **43**, 685-689 (1996).
10. Jurka, J. *Proc. Natl. Acad. Sci.* **94**, 1872-1877 (1997).
11. Jurka, J., Klonowski, P., Trifonov, E. *J. Biomol. Struct. Dyn.* **15**, 717-721 (1998).
12. Feng, Q., Moran, J.V., Kazazian, H.H., Boeke, J.D. *Cell* **87**, 905-916 (1996).
13. Tatout, C., Lavie, L., Deragon J.M. *J. Mol. Evol.* **47**, 463-470 (1998).
14. Ng, S-Y., Ma, J. Jurka, J. (in preparation)

PUBLICATIONS RESULTING FROM THIS WORK

a. Publications in scientific journals (see Appendix)

Jurka, J., Kapitonov, V.V., Klonowski, P., Walichiewicz, J., Smit, A.F.A. Identification of New Medium Reiteration Frequency Repeats in the Genomes of Primates, Rodentia and Lagomorpha. *Genetica* **98**, 235-247 (1996).

Jurka, J., Klonowski, P. Integration of Retroposable Elements in Mammals: Selection of Target Sites. *J. Mol. Evol.* **43**, 685-689 (1996).

Bell, G.I., Jurka, J. The length distribution of perfect dimer repetitive DNS is consistent with its evolution by an unbiased single-step mutation process. *J. Mol. Evol.* **44**, 414-421 (1997).

Jurka, J. Sequence Patterns Indicate an Enzymatic Involvement in Integration of Mammalian Retroposons *Proc. Natl. Acad. Sci.* **94**, 1872-1877 (1997).

Jurka, J., Klonowski, P., Trifonov, E. Mammalian Retroposons Integrate at Kinkable DNA sites. *J. Biomol. Struct. Dyn.* **15**, 717-721 (1998).

Kapitonov, V.V., Jurka, J. MER53, a Non-Autonomous DNA Transposon Associated with a Variety of Functionally Related Defense Genes in the Human Genome. *DNA Sequence. DNA Sequence* **8**, 277-288 (1998).

b. Electronic publications

Rebase Update: <http://www.girinst.org/~server/rebase.html> (1996/1997).

Klonowski, P., Jurka, J. CENSOR Server version 2.0: cursor@charon.girinst.org.

Klonowski, P., Jurka, J. An Annotated Map of Repetitive Sequences in Genbank DNA from Primates: <http://www.girinst.org/public> (1998).

c. Patent (see Appendix)

Jurka, J. Site-directed Recombination. U.S. Patent 5,695,977 (filed in 1995; issued in 1997).