Title:

Surveying Genome to Identify Origins of DNA Replication *In Silico*

Abstract:

DNA replication origins are the bases to realize the process of chromosome replication and analyze the progress of cell cycle.   However, the identification of DNA replication origins localization in whole genome is a labor-intensive task. Here, we propose two novel approaches, ACS and AT, to predict the localization of DNA replication origins *in silico*.   According to our preliminary results, in *Saccharomyces cerevisiae* for example, based on the property that most of DNA replication origins located in intergenic regions, our approaches could cover approximately 60 % data predicted by using other experiment-based approaches, i.e., IBM (Isotope-based Method) (Raghuraman et al. 2001. Science 294:115-121), and ChIP (Chromatin Immunoprecipitation) (Wyrick et al. 2001. Science 294:2357-2360). In accordance with previous papers (Newlon and Theis 2002. Bioessays 24:300-304), two datasets from above two groups show a concordance of approximately 70 %. Surveying genome to identify origins of DNA replication by computational approaches not only provides a faster and predictive tool, but also pre-select targets for further experimentation to molecular biologists.   In the future, we will continue to study the predictivity of DNA replication origins in other species by using our computational approaches, and set up the database of DNA replication origins to be the platform for other research groups to execute the tasks of information analysis, dataset comparison, and data storage.

Keywords:

Authors:

Yang-ren Rau & Huey-jenn Chiang*
Department of Life Science & Institute of Biotechnology, National Dong Hwa University
1, Section 2, Da-Hsueh Road, Shou-Feng
Hualien, Taiwan 97441
R.O.C.

*Corresponding Author
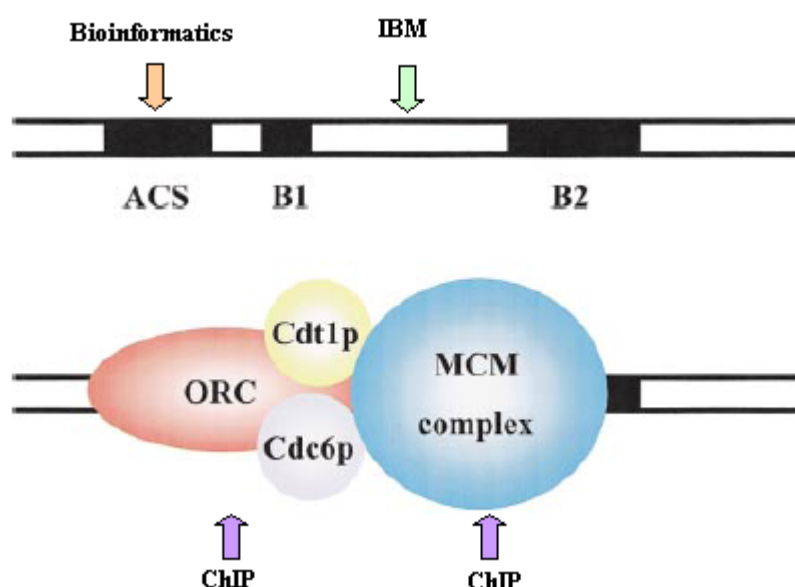E-mail: hjchiang@mail.ndhu.edu.tw; Fax: +886-3-866-2620

Figure:

Figure 1.   Schematic overview of potential origins predicting strategies by using IBM, ChIP, and proposed computational approaches, respectively.   Upper part is the picture shows conserved elements containing in ARS, indicating regions required for the efficient function as a replicating origin.   Lower part is the diagram represents a series of proteins, binding specifically to ARS to function as initializing DNA replication.   The approach of IBM is to screen the locations, the beginning point of DNA replicating, between B1 and B2 elements.   The approach of ChIP is to recognize the positions, the binding sites of ORC (origin recognition complex) and MCM (mini-chromosome maintenance).   ACS approach is to identify the sites of 11-bp ACS [5'-(A/T)TTTA(T/C)(A/G)TTT(A/T)-3'].   AT approach is based on the phenomenon that most DNA replicating origins lie in intergenic regions and contain the AT-rich motifs.

Table:

Table 1.   Length of AT-rich Motifs and Number of AT-rich Motif-Containing Intergenic Regions in *Saccharomyces cerevisiae*.

n	20	21	22	23	24	25	26	27	28	29	30	31	32
Number	905	728	591	499	425	348	295	243	196	167	143	123	113

"n" is the length of AT-rich motifs [5'-(A/T)n-3'] submitted to survey in Saccharomyces intergenic regions in SGD (Saccharomyces Genome Database, http://genome-www.stanford.edu/Saccharomyces/) in Stanford University.
"Number" is the amount of submitted AT-rich motif-containing intergenic regions.

Text:

Introduction:

DNA replication is the process that cells replicate one complete copy of genetic materials before cell division.   Eukaryotic chromosome DNA replication initiates at multiple sites called replication origins.   Most origins of DNA replication in *Saccharomyces cerevisiae* lie in intergenic regions (Raghuraman et al. 2001; Wyrick et al. 2001) and contain an essential 11-base pair (bp) autonomously replicating sequence (ARS) consensus sequence (ACS) (Gilbert 2001; Newlon and Theis 2002; Raghuraman et al. 2001; Wyrick et al. 2001).   The conserved element 11-bp ACS [5'-(A/T)TTTA(T/C)(A/G)TTT(A/T)-3'] is the core of the binding site for the six-subunit initiator protein, origin recognition complex (ORC) (Poloumienko et al. 2001).   Initiation of DNA replication is regulated through the ORC-dependent recruitment of mini-chromosome maintenance (MCM) proteins complex (Wyrick et al. 2001; Newlon and Theis, 2002).   Although most origins contain a perfect match or a one-base mismatch to the ACS, however, the presence of an ACS in genome is not sufficient to predict an origin of replication (Raghuraman et al. 2001; Wyrick et al. 2001).   There are many more ACSs in the genome than origins.

About 200-400 origins of DNA replication are estimated in Saccharomyces genome (Raghuraman et al. 2001; Newlon and Theis 2002).   The plasmid assay for isolation of potential origin (ARS) in *Saccharomyces cerevisiae* was proposed in 1979 (Stinchcomb et al., 1979).   However, before the innovative works to predict the origins in genome-wide scale by using the approaches of IBM (Isotope-based Methods) (Raghuraman et al. 2001) and ChIP (Chromatin Immunoprecipitation) (Wyrick et al. 2001) in 2001, only 67 sites were identified to be the origins of DNA replication in the past 20 years (Wyrick et al. 2001).   The approach of IBM determined the exact origins of replication.   Label DNA with the isotope ^{15}N and ^{13}C (the resulting DNA strands are heavy heavy, HH), and then culture the cells in the medium with two isotopes ^{14}N and ^{12}C.   At various time during S phase, replicated DNA (contains one heavy parent strand and one light daughter strand, HL) was separated from un-replicated DNA (contains two heavy strands, HH) by using density gradient centrifugation.   Hybridization of isolating replicated DNA samples on microarray could recognize the locations of origins.   Totally 332 origins in the genome were identified via this method (at the resolution of about 10 kb).   Another approach named ChIP identified the sites in genome that ORC and MCM proteins bound.   Use the antibodies directed against ORC or MCM, and then the co-precipitated DNA was labeled and hybridized to high-density microarray.   By analyzing the hybridization results, a total of 429 potential origins in genome were

proposed via this method.   Both approaches described above, IBM and ChIP, are set to identify origins of DNA replication on a genome-wide scale, but in different ways. The numbers of origins identified by approaches of IBM (332) and ChIP (429) are consistent with the previous estimates of about 200-400 replication origins in Saccharomyces genome (Raghuraman et al. 2001; Newlon and Theis 2002).   Two datasets show a concordance of approximately 70 % within 10 kb (Newlon and Theis 2002).   Even these two genome-wide approaches are much more efficient than the ones used before; to identity replication origins is a laborious assignment.

We present here two new approaches, ACS and AT, to identify origins of DNA replication in *Saccharomyces cerevisiae* by using computational method.   The ACS approach is based on the knowledge that most origins of DNA replication in *Saccharomyces cerevisiae* contain an essential 11-bp ACS, and lie in intergenic regions.   There are many more ACSs than origins in whole genome (Stillman 2001). However, using 11-bp ACS to search in the "intergenic region," but not "whole genome," might be helpful to predict the locations of potential origins.   Another method, AT approach, is proposed to predict DNA replication origins *in silico* without the necessary of being aware of ACS.   Our novelty is based on the phenomenon that most DNA replicating origins lie in intergenic regions and contain the AT-rich motifs (Mechali, 2001), so the AT-rich motif-containing intergenic regions might be the putative DNA replicating origins.   The performance in identifying origins of DNA replication *in silico* would be presented with the ones predicted by the experiment-based approaches of IBM and ChIP.

Methods:

ACS Approach:

Use 11-bp ACS to search in intergenic regions of *Saccharomyces cerevisiae* in SGD (Saccharomyces Genome Database) (http://genome-www.stanford.edu/Saccharomyces/) in Stanford University.   Due to the order set by SGD, "Pattern Matching" function should be used in the condition when searching for short (<20) nucleotide or peptide sequence.   The sequence of 11-bp ACS [5'-(A/T)TTTA(T/C)(A/G)TTT(A/T)-3'] should be denoted as "WTTTAYRTTTW" for "Pattern Matching" (character "W" means "A/T", character "Y" means "T/C", character "R" means "A/G") to search.   Choose "Not Feature" dataset (includes those portions of the systematic sequence that are not an ORF, centromere, tRNA, RNA gene, or Ty element) for comparing, and set the "Maximum Hits" to "1000," in "Both Strands" option; Set "Mismatch,"   "Deletion," and "Insertion" to "0."   After clicking the icon "Start Pattern Search" and waiting for a

short period of time (in about several minutes), the total hits and matching sequences of ACS, the potential origins, could be retrieved.

nW approach:

Use different length of consecutive (A/T) sequence [(A/T)n= nW, n= length] to search in intergenic regions of *Saccharomyces cerevisiae* in SGD (Saccharomyces Genome Database) (http://genome-www.stanford.edu/Saccharomyces/).   According to the basic operations described in previous Method "ACS Approach", count the amount of intergenic sequences containing desired length of consecutive (A/T) sequence.   Then, decide the expected amount of replication origins, in accordance with the number determined by microscopic studies, theoretical calculations, or other experience-based approaches, etc.   At last, choose the amount of intergenic sequences close to the expected one.

Results and Discussion:

Amount of Origins Predicted by Using ACS and nW Approaches:

Figure 1 is the schematic overview of potential origins predicting strategies by using IBM, ChIP, and bioinformatics approaches.   The approach of IBM is to screen the beginning point of DNA replicating.   The OBR (origin of bidirectional replication) is between B1 and B2 elements (Gerbi and Bielinsky, 1997).   The approach of ChIP is to recognize the positions that the ORC (origin recognition complex) and MCM (mini-chromosome maintenance) bind.   Our proposed ACS approach is using 11-bp ACS to search in the "intergenic region."   Totally, there are 466 ACS-containing intergenic regions, the putative DNA replication origins, to be found via ACS approach.   Besides, table 1 is the result of nW approach: length of AT-rich motifs and number of AT-rich motif-containing intergenic regions in *Saccharomyces cerevisiae*. We choose the dataset whose amount of AT-rich motif-containing intergenic regions as 499, since about 400 origins were estimated, 332 by IBM, 429 by ChIP and 466 by ACS approaches.

Results Comparison by Using Different Approaches:

Comparison of the results of IBM and ChIP shows good overlapping: nearly a quarter were separated by less than 1 kb, half by less than 3 kb, and two-thirds by less than 5 kb, and approximately 70 % within 10 kb (Newlon and Theis, 2002).   By the same criterion, compare the result of IBM to the one of ACS approach: 10.24 % within 1 kb, 25.60 % within 3 kb, 37.35 % within 5 kb, and 59.04 % within 10 kb; Compare the result of ChIP to the one of ACS approach: 31.00 % within 1 kb, 39.16 % within 3 kb, 45.22 % within 5 kb, and 61.54 % within 10 kb.   It demonstrates good matches that

potential origins predicted by using ACS approach with the ones by using IBM and ChIP.

Although DNA replication origins predicted by ACS approach, using 11-bp ACS to survey in intergenic region, could cover 60 % sites predicted by both approaches of IBM and ChIP in the case of *Saccharomyces cerevisiae*, no consensus sequences were identified as necessary and important in other eukaryotes as the 11-bp ACS in Saccharomyces until now (Mechali, 2001).   It is generally accepted that DNA replication initiates at specific sites, but not clear whether specific consensus sequence elements are used to specify these origins, or whether higher levels of regulation are more important.

By the same criteria, comparison of results of IBM and AT approach shows coverage of 10.54 % within 1 kb, 25.60 % within 3 kb, 36.75 % within 5 kb, and 59.94 % within 10 kb; comparison of results of ChIP and AT approach shows coverage of 13.29 % within 1 kb, 23.31 % within 3 kb, 33.57 % within 5 kb, and 54.31 % within 10 kb.   This demonstrates good overlapping of putative origins predicted by using AT approach with the ones by utilizing *in vivo* experiment-based methods.   It can be executed without being aware of specific consensus sequence, and can be applicable as long as the target species have their whole genome sequenced.   If the AT approach proposed in this study could provide the correct identification of more than half (close to 60 %) of the origins by using IBM and ChIP (corresponding between IBM and ChIP is 70 %) in other cases, just like the example demonstrating in *Saccharomyces cerevisiae* here, it would provide a good complementary tool to the other experiment-based approaches for screening the putative origins and pre-select targets for further experimentation to molecular biologists.

Since the new computational approach based on AT-rich motif-containing intergenic regions could work without information of ACS, it could predict the localization of origins *in silico* as long as the target species have their whole genome sequenced. Until now (2003/3), according to the "Eukaryotae Genomes Taxonomy / List" information in NCBI (http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/euk_g.html), there are nine eukaryotae species have whole genome sequenced: *Anopheles gambiae*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Encephalitozoon cuniculi*, *Guillardia theta nucleomorph*, *Saccharomyces cerevisiae*, *Plasmodium falciparum*, and *Schizosaccharomyces pombe*.

Besides the work of *Saccharomyces cerevisiae* has been done here to predict the

localization of DNA replication origins and compare the dataset with the genome-wide experiment based prediction, there are eight other eukaryotic species can be studied (and the number of eukaryotic species whose genome has been sequenced will increase by the progressing of genome sequencing projects).

DNA replication origins are fundamental to realize the process of chromosome replication and the progress of cell cycle, but the identification of DNA replication origins localization in whole genome is a laborious job.   The proposed computational approach can provide alternative strategy complementary to the genome-wide microarray-based analysis and pre-select targets for further experimentation to molecular biologists.

References:
Gerbi, S. A. and Bielinsky, A. K. 1997. Replication initiation point mapping. Methods 13:271-280.

Gilbert, D. M. 2001. Making sense of eukaryotic DNA replication origins. Science 294:96-100.

Newlon, C. S. and Theis, J. F. 2002. DNA replication joins the revolution: whole-genome views of DNA replication in budding yeast. Bioessays 24:300-304.

Poloumienko, A., Dershowitz, A., De, J., and Newlon, C. S. 2001. Completion of replication map of *Saccharomyces cerevisiae* chromosome III. Mol. Biol. Cell. 12:3317-3327.

Raghuraman, M. K., Winzeler, E. A., Collingwood, D., Hunt, S., Wodicka, L., Conway, A., Lockhart, D. J., Davis, R. W., Brewer, B. J., and Fangman, W. L. 2001. Replication dynamics of the yeast genome. Science 294:115-121.

Stillman, B. 2001. DNA replication. Genomic views of genome duplication. Science 294:2301-2304.

Stinchcomb, D. T., Struhl, K., and Davis, R. W. 1979. Isolation and characterisation of a yeast chromosomal replicator. Nature 282:39-43.

Wyrick, J. J., Aparicio, J. G., Chen, T., Barnett, J. D., Jennings, E. G., Young, R. A., Bell, S. P., and Aparicio, O. M. 2001. Genome-wide distribution of ORC and MCM

proteins in *S. cerevisiae*: high-resolution mapping of replication origins. Science 294:2357-2360.