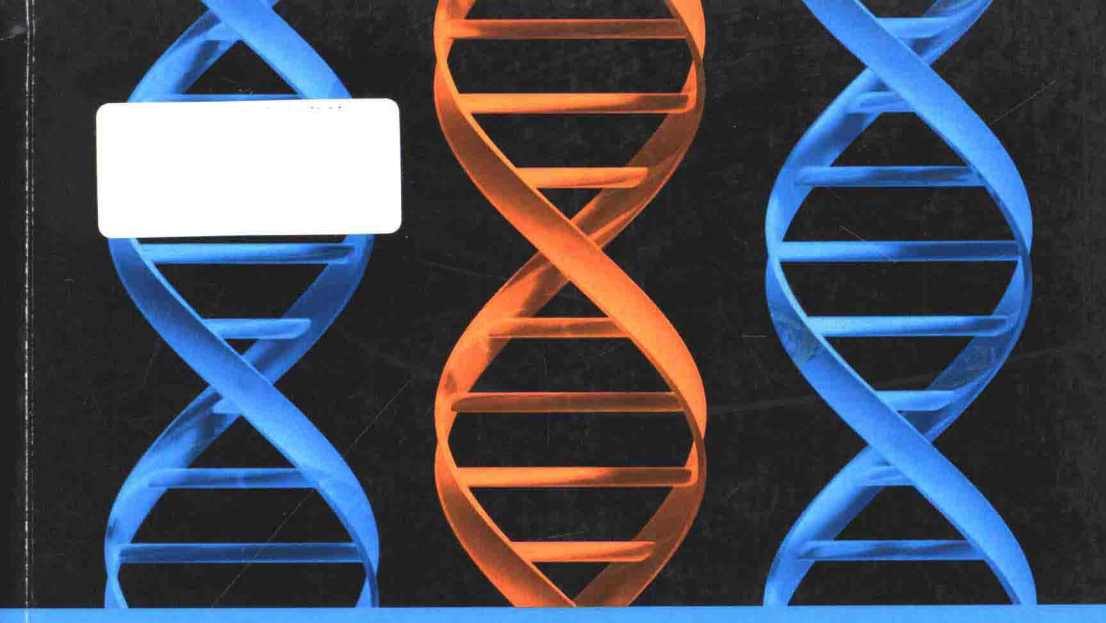

Huaichun Wang

The Effects of Nucleotide Bias on Genome Evolution

The causes and effects of wide variations in G+C content of the genomes

VDM
Verlag
Dr. Müller

Huaichun Wang

The Effects of Nucleotide Bias on Genome Evolution

The causes and effects of wide variations in G+C content of the genomes

VDM Verlag Dr. Müller

Impressum/Imprint (nur für Deutschland/ only for Germany)
Bibliografische Information der Deutschen Nationalbibliothek: Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Coverbild: www.purestockx.com

Verlag: VDM Verlag Dr. Müller Aktiengesellschaft & Co. KG
Dudweiler Landstr. 99, 66123 Saarbrücken, Deutschland
Telefon +49 681 9100-698, Telefax +49 681 9100-988, Email: info@vdm-verlag.de
Zugl.: Ottawa, University of Ottawa, PhD dissertation, 2005

Herstellung in Deutschland:
Schaltungsdienst Lange o.H.G., Berlin
Books on Demand GmbH, Norderstedt
Reha GmbH, Saarbrücken
Amazon Distribution GmbH, Leipzig
ISBN: 978-3-639-19164-6

Imprint (only for USA, GB)
Bibliographic information published by the Deutsche Nationalbibliothek: The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de .

Cover image: www.purestockx.com

Publisher:
VDM Verlag Dr. Müller Aktiengesellschaft & Co. KG
Dudweiler Landstr. 99, 66123 Saarbrücken, Germany
Phone +49 681 9100-698, Fax +49 681 9100-988, Email: info@vdm-publishing.com

Printed in the U.S.A.
Printed in the U.K. by (see last page)
ISBN: 978-3-639-19164-6

Huaichun Wang

The Effects of Nucleotide Bias on Genome Evolution

To the memory of my father

Acknowledgements

I would like to thank my supervisor, Professor Donal Hickey for assigning me this interesting thesis topic and directing every detail of the whole thesis work, and lots of financial support for attending conferences and summer schools. I also thank the members of my advisory committee, Drs. George Carmody, Guy Drouin and Marcel Turcotte for their advice. Dr. Xuhua Xia also gave very insightful advice that is greatly appreciated. Ada Chyurlia always gave me help in need and Dr. Greg Singer helped a lot by maintaining the Sun server (even after he had left the lab), which was essential for my programming. Financial supports from The Natural Sciences and Engineering Research Council (NSERC) of Canada (to Dr. Hickey), Ontario Graduate Scholarship and University of Ottawa graduate scholarships are appreciated. Finally, as a student at a senior age, family supports are all important. I would like to thank my parents for bringing me up and my brothers and sister for taking care of them so I can study uninterrupted for the past three and a half years. My wife Shuli and my son Sean gave me love and joy so that my study was less stressful.

Abstract

The genomic G+C content of prokaryotes varies from approximately 23% to 77% among genomes. In contrast, among vertebrates, the variation is greatest within the same genome rather than between genomes. There has been a long-standing controversy concerning the causes of these inter- and intra-specific variations. Is it caused by natural selection, favored by the selectionists or, conversely, is it selectively neutral (the neutralist view)? In this study, we investigated the source of nucleotide compositional variation (nucleotide bias) and the consequences of the bias on protein sequence and genome evolution. Thermal adaptation is a primary example to study the effect of natural selection and has been thoroughly studied in this project. We found that both GC content and length of ribosomal RNA genes show positive correlations with optimal growth temperature in prokaryotes and these correlations are not due to phylogenetic history. The correlations are concentrated almost entirely within the stem regions of the rRNA. The rRNA loops, however, show very constant base composition regardless of temperature optima or genomic GC content. The loops were found to have very high amount of adenosine nucleotides throughout prokaryotes and eukaryotes. These results clearly demonstrated that environmental temperature is a selective force that drives rRNA gene evolution and different segments of the same gene (i.e., the stems and loops of the rRNA gene) experience differential selection, although the mutation spectrum presumably should be similar between the loops and stems.

For protein coding genes, mutation and natural selection play a different role compared to the rRNA genes. The neutralist predicts mutational bias would cause protein sequence evolution, while the selectionist would predict that the protein sequence is not related to genomic GC content. To investigate these two postulations and analyze the consequences of nucleotide bias in eukaryotic genomes, we studied homologous genes and their encoded proteins in two flowering plants, *Oryza sativa* (rice) and *Arabidopsis thaliana*. While there is a relatively homogenous GC content in the *Arabidopsis* genes (26% to 69%), the GC content of the rice genes is very heterogeneous (27% to 90%). High GC rice genes encode proteins having a high frequency of GC-rich codons encoded amino acids, *i.e.*, glycine, alanine, arginine and proline. Low GC rice genes and

Arabidopsis genes encode proteins having a high frequency of AT-rich codons encoded amino acids, *i.e.*, phenylalanine, tyrosine, methionine, isoleucine, asparagines and lysine. Furthermore, the effects of nucleotide bias on synonymous codon usage in the rice and *Arabidopsis* genomes were studied. We have shown that synonymous codon usage in the rice genome is primarily dictated by the GC content of the genes, rather than by translational selection. This study in multicellular higher plants, together with previous work on prokaryote and yeast, provide persuasive evidence that mutational nucleotide bias is a cause, rather than a consequence, of protein evolution and this affects codon usage and protein composition in a predictable way.

Résumé

Le contenu génomique en nucléotides G+C des prokaryotes varie approximativement de 23% à 77% entre génomes. Chez les vertébrés, contrairement aux prokaryotes, cette variation est plus élevée au sein d'un génome plutôt qu'entre génomes différents. Il existe une controverse de longue durée concernant les causes de ces variations inter- et intra-spécifique. Est-ce que ces variations sont le résultat de la sélection naturelle, cette explication est favorisée par les sélectionnistes ou d'une sélection neutre (point vue des neutralistes)? Dans cette étude, nous examinons la source des variations de la composition en nucléotides (biais nucléotidique) et les conséquences de ce biais sur les séquences de protéines et sur l'évolution du génome. L'adaptation thermique convient bien à l'étude de l'effet de la sélection naturelle et nous l'avons étudié de façon rigoureuse dans ce projet. Nous avons trouvé que le contenu en nucléotides GC et la longueur des gènes d'ARN ribosomal (ARNr) montrent une corrélation positive avec les températures optimales de croissance chez les prokaryotes et que ces corrélations ne sont pas le résultat de l'histoire phylogénétique des espèces. Les corrélations sont concentrées principalement au niveau des tiges des ARNr. Par contre, les bras d'ARNr sont très constants dans leur composition en nucléotides et ne sont pas affectés par les températures optimales ou le contenu en nucléotides GC. Il semblerait que les bras abondent de nucléotides d'adénosines et ce autant chez les procaryotes que chez les eucaryotes. Ces résultats démontrent clairement que la température environnementale exerce une force sélective qui conduit à l'évolution des gènes ARNr et que différents segments du même gène (i.e., les tiges et les bras du gène d'ARNr) sont affecté par différente sélection même si le spectre des mutations est présumé être similaire dans les tiges et les bras.

Pour les gènes codant pour des protéines, les mutations et la sélection naturelle jouent un rôle différent comparé aux gènes ARNr. Les neutralistes prédissent qu'un biais mutationnelle chez les séquences protéiques n'est pas associé au contenu en GC génomique. L'investigation de ces deux postulats et analyser les conséquences du biais nucléotidique, nous avons étudie des gènes homologues et qui codent pour des protéines de deux plantes à fleurs: *Oryza sativa* (riz) et *Arabidopsis thaliana*. Pendant que le contenu en GC est relativement homogène dans les gènes d'*Arabidopsis* (26% à 69%), le

contenu en GC chez le riz est très hétérogène (27% à 90%). Les gènes de riz à haute teneur en GC codent pour des protéines dont la fréquence d'acides aminés ayant des codons riches en nucléotides GC est élevée, i.e. glycine, alanine, arginine et proline. Chez le riz et *Arabidopsis*, les gènes à plus faible teneur en GC codent pour des protéines composé d'acide aminés ayant des codons riches en base AT, i.e, phenylalanine, tyrosine, methionine, isoleucine, asparagines et lysine. De plus, les effets du biais nucléotidique sur l'utilisation des codons synonymes du génome du riz est principalement contrôlé par le contenu en nucléotides GC des gènes, plutôt que par la sélection traductionelle. Cette étude des plantes multicellulaires d'ordre supérieur, de même que des recherches passées sur les prokaryotes et levures, démontrent de façon claire et précise que le biais nucléotidique est une cause plutôt qu'une conséquence de l'évolution des protéines et ceci affecte l'utilisation des codons et la composition protéique que nous pouvons prédire.

RC: (codon) redundancy class

rRNA: ribosomal RNA

RSCU: Relative Synonymous Codon Usage

ssu rRNA: small subunit rRNA

Thermo: thermophiles (thermophilic species)

tRNA: transfer RNA

TT: thymine dimer

UV: ultraviolet

VSP: (DNA repair) very short patch

IUPAC code table

Amino acid codes

1-letter	3-letter	description
A	Ala	Alanine
R	Arg	Arginine
N	Asn	Asparagine
D	Asp	Aspartic acid
C	Cys	Cysteine
Q	Gln	Glutamine
E	Glu	Glutamic acid
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
L	Leu	Leucine
K	Lys	Lysine
M	Met	Methionine
F	Phe	Phenylalanine
P	Pro	Proline
S	Ser	Serine
T	Thr	Threonine
W	Trp	Tryptophan
Y	Tyr	Tyrosine
V	Val	Valine
B	Asx	Aspartic acid or Asparagine
Z	Glx	Glutamine or Glutamic acid

Nucleic acid codes

code	description
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
U	Uracil
R	Purine (A or G)
Y	Pyrimidine (C, T, or U)
M	C or A
K	T, U, or G
W	T, U, or A
S	C or G
B	C, T, U, or G (not A)
D	A, T, U, or G (not C)
H	A, T, U, or C (not G)
V	A, C, or G (not T, not U)
N	Any base (A, C, G, T, or U)

Table of Contents

List of Tables