The Implicit Genome

Lynn Helena Caporale,
Editor

OXFORD UNIVERISITY PRESS

The Implicit Genome

This page intentionally left blank

The Implicit Genome

Edited by
Lynn Helena Caporale

OXFORD
UNIVERSITY PRESS
2006

OXFORD
UNIVERSITY PRESS

Oxford University Press, Inc., publishes works that further
Oxford University's objective of excellence
in research, scholarship, and education.

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Published by Oxford University Press, Inc.
198 Madison Avenue, New York, New York 10016

www.oup.com

Oxford is a registered trademark of Oxford University Press

Library of Congress Cataloging-in-Publication Data

The implicit genome / edited by Lynn Helena Caporale.
p. ; cm.
Includes bibliographical references and index.
ISBN-13: 978-0-19-517270-6; 978-0-19-517271-3 (pbk).
ISBN-10: 0-19-517270-1; 0-19-517271-X (pbk).
1. Evolutionary genetics.
[DNLM: 1. Genome. 2. Evolution, Molecular. 3. Immunity—genetics.
4. Repetitive Sequences, Nucleic Acid—physiology. QH 447 I34 2006] I. Caporale, Lynn Helena.
QH390. I47 2006
572.8′38—dc22 2005011590

9 8 7 6 5 4 3 2 1

Printed in the United States of America
on acid-free paper

To Mom and Dad

and to the memory of my Grandparents

This page intentionally left blank

Contents

Contributors

Lynn Helena Caporale (Editor)
The Judith P. Sulzberger MD Genome Center, Columbia University, 1150 St Nicholas Avenue, New York, NY 10032

Alyson Ashe
School of Molecular and Microbial Biosciences, Biochemistry Building—G08, University of Sydney, New South Wales 2006, Australia

J. David Barry
Wellcome Centre for Molecular Parasitology, University of Glasgow, Anderson College, 56 Dumbarton Rd, Glasgow G11 6NU, United Kingdom

Christopher D. Bayliss
Oxford University Molecular Infectious Diseases Group, Weatherall Institute for Molecular Medicine, John Radcliffe Hospital, Headington, Oxford OX3 9DU, United Kingdom

Rupert Beale
Medical Research Council Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, United Kingdom

Rhona H. Borts
Department of Genetics, University of Leicester, University Road, Leicester LE1 7RH, United Kingdom

Marie Csete
Emory University School of Medicine, 1462 Clifton Rd NE, Room 420, Atlanta GA 30322

Donald M. Crothers
Department of Chemistry, Yale University, 350 Edwards Street, PO Box 208107, New Haven, CT 06520-8107

John Doyle
California Institute of Technology, CDS 107-81, 1200 E California Blvd, Pasadena, CA 91125-8100

Claire Fraser
The Institute for Genome Research, 9712 Medical Center Drive, Rockville, MD 20850

Errol C. Friedberg
Laboratory of Molecular Pathology, Department of Pathology, University of Texas Southwestern Medical Center at Dallas, Dallas, TX 75390

Brenton R. Graveley
Department of Genetics and Developmental Biology, University of Connecticut Health Center, 263 Farmington Avenue, Farmington, CT 06030-3301

Ellen Hsu
Department of Physiology and Pharmacology, State University of New York Health Science Center, Brooklyn, NY 11203

Dagmar Iber
Mathematical Institute, Centre for Mathematical Biology, St John's College, University of Oxford, 24–29 St Giles, Oxford OX1 3LB, United Kingdom

Carolyn L. Jahn
Department of Cell and Molecular Biology, Feinberg School of Medicine, Northwestern University, 303 East Chicago Ave., Chicago, IL 60611

Yechezkel Kashi
Department of Biotechnology and Food Engineering, Technicon-Israel Institute of Technology, Haifa 32000, Israel

David G. King
Department of Zoology, Southern Illinois University, Carbondale, IL 62901-6899

David T. Kirkpatrick
Department of Genetics, Cell Biology and Development, University of Minnesota, 6–160 Jackson Hall, 321 Church St SE, Minneapolis, MN 55455

Nancy Maizels
Departments of Immunology and Biochemistry, University of Washington Medical School, 1959 N.E. Pacific Street, Seattle, WA 98195-7650

E. Richard Moxon
Oxford University, Molecular Infectious Diseases Group, Weatherall Institute for Molecular Medicine, John Radcliffe Hospital, Headington, Oxford OX3 9DU, United Kingdom

Garry S. A. Myers
The Institute for Genome Research, 9712 Medical Center Drive, Rockville, MD 20850

Ian T. Paulsen
The Institute for Genome Research, 9712 Medical Center Drive, Rockville, MD 20850

Eduardo P. C. Rocha
Unité Génétique des Génomes Bactériens, Institut Pasteur, 28 rue du Dr. Roux, 75724 Paris Cedex 15, France and Atelier de Bioinformatique, Université Pierre et Marie Curie, 12, Rue Cuvier, 75005 Paris, France

Harold C. Smith
Department of Biochemistry and Biophysics, Box 712, University of Rochester, School of Medicine and Dentistry, 601 Elmwood Avenue, Rochester, NY 14642

Edward N. Trifonov
Genome Diversity Center, Institute of Evolution, University of Haifa, Mount Carmel, Haifa 31905, Israel

Susan R. Wessler
University of Georgia, Department of Plant Biology, Miller Plant Sciences Building, Athens, GA 30602

Emma Whitelaw
School of Molecular and Microbial Biosciences, Biochemistry Building—G08, University of Sydney, New South Wales 2006, Australia

This page intentionally left blank

The Implicit Genome

This page intentionally left blank

An Overview of the Implicit Genome

Lynn Helena Caporale

I was in a new world, and … could not help speculating on what my wanderings there … might bring to light

Alfred Russel Wallace, *The Malay Archipelago* (1869)

Most analyses assume that genomes are to be read as linear text, much as a sequence of nucleotides can be translated into a sequence of amino acids by looking in a table. However, information can evolve in genomes with distinct forms of representation, such as in the structure of DNA or RNA and the relationship between nucleotide sequences. Such information has importance to biology yet is largely unexpected and unexplored. As described in this volume, much of this information, through mechanisms ranging from alternative splicing of RNA to the generation of bacterial coat protein diversity, affects the probability of distinct types of alterations in the nucleic acid sequence. Some genomic DNA sequences affect genome stability, handling, and organization, with implications for the robustness of lineages over evolutionary time. The examples reviewed in this volume, taken from a broad range of biological organisms, both extend our view of the nature of information encoded within genomes and deepen our appreciation of the power of natural selection, through which this information, in its various forms, has emerged.

Introduction

There was a glorious moment, in the 1960s, when, like children first learning to read, we began to perceive meaning in strings of nucleotides in DNA. Suddenly, we could understand that TTT in a protein-coding sequence of DNA meant that the amino acid phenylalanine would be incorporated into a protein. We could, in our minds, and later with computers, directly translate strings of bases,[1] taken as triplets,[2] into strings of amino acids, which then fold up to form three-dimensional proteins.

While the challenge remained of perceiving the three-dimensional structure specified by a linear protein sequence, we understood that information was encoded in DNA in a way that was both explicit and linear.

But now, with complete genome sequences before us, instead of triumph comes humility. The strings of nucleotides that we proudly translate into protein amount to less than 2% of the human genome.[3] We know well that additional DNA sequences are involved in the regulation of expression.[4,5] However, fully understanding the information content of genomes will involve expanding our imagination with respect to both what types of information may be there and how information might be represented.

Novel forms of information may be represented in the structure of DNA or in the relationships between sequences, rather than in the sequences themselves. This book focuses on information that leads to editing, splicing, recombination, mobility, and/or which affects the rate, type, and location of mutations of nucleic acid sequences. Such variations in the probability of distinct changes in DNA sequence have implications for evolution. In fact, for many specific examples in this book, certain types of mutation, with an increased probability of being adaptive, are more probable than would be expected from the "average" mutation rate. The reader is challenged to decide whether these examples are just isolated, albeit interesting, cases, specific, for example, to the stress of host–pathogen interactions, which provide no broader insight into genome organization and evolution; alternatively, the reader may decide that the examples of focused genetic variation discussed in this book represent the tip of an iceberg that should command more immediate attention.

One Nucleotide Sequence Can Imply Other Nucleotide Sequences

We recognize information that is encoded in DNA explicitly: in the right context and reading frame, TTTGGG encodes the sequence phenylalanine-glycine. But as the chapters in this book make clear, even information that we have been confident about our ability to interpret, such as protein-coding information, often only is implied in a genome.

Editing and Splicing

Due to splicing and editing of nucleic acid sequences, genomic DNA sequences often do not directly match the sequences of the proteins encoded by the genome.

Alternative Splicing

Although stunning when first observed, we now tend to view the concept of "genes in pieces" as routine—until we attempt to identify the proteins encoded by a newly-sequenced eukaryotic genome. Chapter 15 describes the *Drosophila Dscam* gene, a single gene that, at least theoretically, can, by alternative splicing, generate three times as many proteins as there are protein-coding genes in the genome. For

RNA transcripts to be spliced, they must contain additional information, beyond the base triplets that encode amino acids. This information, which is located both within sequences destined to be exons and those destined to be introns, and across their boundaries, implies the resulting context-dependent pattern of splicing and the "either/or" relationships between certain exons. Exon choice that feeds back to be self-perpetuating can serve as a developmental switch, most famously in sex determination in *Drosophila*.

RNA Editing

When "translating" nucleic acid sequences, we expect to be able to predict the protein sequence, based solely on the table of codons. However, as described in chapter 14, it is not unusual for the primary RNA transcript to be edited as well as spliced before it is translated into protein, so that the RNA sequence no longer is complementary to the DNA sequence from which it was transcribed. Before editing was understood, comparisons between protein and DNA sequences in plant organelles made it appear that chloroplasts use a different triplet code.

In an example cited in chapter 14, at a position that corresponds to an arginine in the protein, the genomic sequence of a human calcium channel is CAG, which bioinformatics software would translate as glutamine; editing to a codon that is translated as arginine depends upon a relationship between DNA sequences: complementarity between the exon that contains the CAG and its neighboring intron.

While "changing" glutamine to arginine (which has an important effect on the properties of the ion channel) requires alteration of a single base in the message, the so-called "cryptogenes" of kinetoplastids undergo far more extensive editing. Some phenylalanines are encoded entirely by Us that are added when the RNA transcript is edited. For software to find in a protein database the protein sequence encoded by that stretch of DNA, it would have to be designed based on an understanding of how the primary transcript is edited upon interaction with separately encoded "guide RNAs."

DNA Gymnastics

As a routine part of their life cycle, ciliates alter their DNA dramatically. As described in chapter 13, although ciliates are unicellular they have two "genomes," each in a separate DNA-containing organelle. The nucleus that undergoes meiosis contains the "germline" genome, while genes are transcribed into mRNA using the other nucleus, which contains the "somatic" genome. In forming their somatic genome, some ciliates eliminate up to 95% of the germline genome sequences, then rearrange and amplify the remainder to generate as many as 1000 copies of each of their genes.

After meiosis and exchange of haploid nuclei with another individual to form the zygote, ciliates must define the sequences that will become part of the new somatic nucleus. Because for some ciliates the order of nucleotides in the germline does not correspond to the order of amino acids in their proteins, germline sequences must also be unscrambled. As described in chapter 13, in the equivalent

of a subtractive hybridization experiment, double-stranded RNA moves from the meiotic "germline nucleus" to the old somatic nucleus and then to the developing new somatic nucleus; dsRNA that enters the new somatic nucleus is "interpreted" as a sequence that was not found in the old somatic nucleus, and thus should be deleted. Chapter 13 points out that analysis of how information flow is handled in organisms operating with distinct nuclei could provide a window into DNA sequence "requirements for the divergent nuclear and genomic functions of transcription and partitioning of genetic material."

Because, as described in chapter 13, siRNA and heterochromatin appear to be involved, deletion can be considered an extreme case of "turning off" regions of DNA. In addition, because the sequences that are eliminated between the germline and somatic nuclei have homology with transposable elements, ciliates may represent an extreme example of the potential for transposable elements to define special contexts or behaviors in a genome.

As chapter 13 points out, from telomeres to RNA splicing, to the role of histone acetylation in transcription activation, we have learned a lot by studying ciliates; it appears likely we will learn more, about genome organization and meiosis. The concept that the somatic and germline nuclei communicate via siRNA is intriguing. Our perspective on the biochemical processes that occur as information is transferred between generations likely would be quite different if decades ago molecular geneticists had selected a ciliate, such as *Oxytricha nova,* rather than yeast, as a model unicellular eukaryote organism.[6]

Rules Rather Than Genes

The numbers of different antibodies that can be generated by the vertebrate immune system, as described in chapter 9, and different surface antigens that can be generated by the pathogens that afflict us, as described in chapter 5, are orders of magnitude greater than the number of genes encoded in their respective genomes. Yet if you search the human germline genome for intact genes encoding antibodies, you will not find them there. Both vertebrates and pathogen genomes imply, rather than specify, this diversity of protein sequences, by storing information in gene fragments, along with rules for their assembly. (Alternative splicing as described in chapter 15 provides another way to store information for multiple related protein sequences efficiently.)

The Immunoglobulins

The initial step in construction of a gene encoding an immunoglobulin heavy chain is the recombination of one of a palette of genomic variable regions beside another gene segment. As detailed in chapter 9, information in the flanking "recombination signal sequences" defines which classes of potentially functional gene segments can recombine with each other. In one startling aspect of the formation of an immunoglobulin heavy-chain gene, "diversity" regions may be inserted in any of their three reading frames in a combinatorial assortment next to "variable" regions. In addition, some diversity regions can be inserted in two orientations (i.e. so that

either strand can be the "coding" strand), such that the same small patch of DNA may generate six different amino acid sequences. As emphasized in chapter 9, diversity regions also can vary in length, and nontemplated sequences may be added at the junction; this further increases the unpredictable (from the point of view of pathogens) diversity of immunoglobulin binding sites. Thus, the precisely targeted diversity of the immunoglobulins is implied in the genomic DNA.

The Pathogens

Trypanosome surface antigen diversity comes from a repertoire of DNA sequences embedded in cassettes. Flanking sequences facilitate replacement of an antigen that is in an expression site with a copy of another, hopefully (from the pathogen's perspective) not cross-reacting, antigen from the cassette archive.

In addition, as an infection proceeds, mosaic surface antigens may be assembled from multiple partial sequences drawn from an extensive repertoire of "pseudogenes." Chapter 5 points out that generation of mosaic antigens from pseudogenes makes tens of millions, or even billions, of potential surface variants theoretically possible.

As described in chapter 5, diverse pathogens, including *Neisseria*, *Pneumocystis carinii*, and *Borrelia burgdorferi*, "hide" from their "hosts" by varying their surface antigens rapidly during the course of infection by generating mosaic antigens via gene conversion. Enzymes that catalyze recombination at specific target sequences generate antigenic diversity by such mechanisms as inverting a promoter involved in the expression of surface proteins or replacing a patch of DNA sequence (as described in chapter 7).

Mutable Sequences: Adjustments and Reversible Inactivation

When studying the complex mechanisms that regulate the expression of genes, we traditionally have assumed that a gene's intrinsic activity is arrived at by the action of selection on multiple random point mutations across evolutionary time. However, several chapters in this book, which use terms such as "tuning knobs" (mutable repeats: chapter 4) and "rheostats" (chapter 8), describe biochemical mechanisms that can expedite intergenerational exploration and adjustment of the range of gene activity through less damaging routes than random genome-wide point mutation.

Diversity

Chapters 3, 4, 7, and 11 describe sequences that mutate at frequencies that can be greater than 1000-fold higher than the genome average. Tandem repetitive sequences, such as GGGGGGGGG or CAGCAGCAG, change length frequently; thus such sequences imply a tendency to generate diversity at that locus between generations. Such highly mutable and statistically improbable repetitive sequences are overrepresented in bacterial and eukaryotic genomes but are underrepresented in

conserved regions of constitutively-expressed housekeeping genes. Evidence, discussed in chapters 3 and 4, that these mutable sequences are under selection indicates that their mutability affects fitness. Within a bacterial population and also among the descendants of an individual bacterium the rapid rate of mutation at repeat sequences generates diversity at loci, such as those for surface antigens, at which diversity has provided a selective advantages in the past.

Chapter 3 points out that with just 12 genes able to switch between "on" and "off" due to reversibly mutable repeat sequences, a population of bacteria descended from a single individual can explore 4096 (2^{12}) different "versions" of its genome without damage to other loci en route. Thus, when we obtain one complete genome sequence we see only an example of the genomic range of that species. Chapter 3 introduces the concept of a "species genome," which takes into account the "full repertoire of variations in repetitive DNA tracts and locations of insertion elements." The diverse phenotypes of the full set of genomes that are implied in the repetitive sequences of an individual genome extend the potential adaptive capabilities of its descendants to meet classes of challenges that can in effect be anticipated based upon the species' past experience.

As suggested in chapter 4, recent data—ranging from *Drosophila* to cattle—hint that mutable sequences not only can differ among cells in an individual but also are involved in rapid phenotypic adjustments in eukaryotes. Repetitive microsatellites in the human genome can have meiotic mutation rates as high as several percent per generation. These highly mutable loci contribute so much diversity to human populations that some are used for DNA "fingerprinting."

Adjustments

Mutable repeats have the potential not only for generating diversity within a population but also for facilitating the reversible "adjustment" of activity, through selected changes in the frequency of these highly polymorphic alleles, at multiple loci, from generation to generation. In other words, in modeling evolution it is important to consider not only diversity in a contemporary population but also the potential for distinct classes of diversity in the population descended from an individual genome.

Changes in the length of amino acid repeats (encoded by triplet repeats at the DNA level) can have the effect of adjusting biochemical properties such as protein flexibility, affinity for substrate, and strength of protein–protein interactions. As chapter 4 points out, such repeats are markedly overrepresented in transcription factors, protein kinases, and genes encoding developmental regulatory proteins.

As described in chapter 3, changes in length of repetitive sequences located between the right and left sides of the promoter affect the level of transcription in prokaryotes. Variations in the number of repeats can affect transcription activity in eukaryotes as well. Chapter 4 provides the intriguing example of a polymorphic repeat in the first intron of a rate-limiting enzyme in the synthesis of catecholamine neurotransmitters, and estimates that such hypermutable loci may be found in the regulatory regions of as many as several thousand human genes.

As discussed in chapter 6, overrepresentation of closely spaced repeats in mismatch repair genes results in a tendency for recombination to reduce or eliminate the activity of these genes in a proportion of the bacterial population (activity can be recovered through recombination with DNA taken up from the environment). Those bacteria with decreased mismatch repair activity have an increased rate of generation of diversity at repeats elsewhere in the genome that become more unstable when mismatch repair is decreased.

Chapter 8 suggests that the level of transcription also can be affected by the presence of transposable elements in introns. Chapter 14 suggests that the extent of editing of multisubunit ion channel mRNA can serve as a "rheostat", a term also used in chapter 16 with reference to imprinting fo dosage-sensitive genes. As described in chapter 16, epigenetic regulation can adjust, during development, the level of expression of certain alleles through silencing of neither, one, or both alleles in different cells. All may facilitate the combination of robustness and stability that modularity can contribute to development.[7,8]

Forms of Information

Multiple Levels of Messages: Using the Degeneracy of the Genetic Code

Because more than one codon is available to specify most amino acids, additional information may be transmitted along with a protein-coding sequence. Such additional information will constrain the choice between what are otherwise considered to be synonymous codons.[9]

For example, chapter 6 presents two lines of evidence that suggest that selection for increased mutability leads to codon choices that result in the overrepresentation of closely spaced repeats in mismatch repair genes. First, due to the degeneracy of the genetic code, the same sequence of amino acids can be encoded by any one of many less mutable DNA sequences. Second, chapters 4 and 6 present examples in which the presence of a repeated sequence, and thus the property of hypermutability, was conserved, while the nature of the repeat and the amino acid sequence it encoded was not conserved.

In *B. burgdorferi*, the requirement for a conserved nucleotide repeat to enable antigenic variation of a surface protein by gene conversion between two repeats of the five amino acids EGAIK constrains, to a single DNA sequence, what otherwise would be a choice among nearly 200 synonymous sequences that could encode each EGAIK.

As described in chapter 10, information that affects the tendency to hypermutate along a DNA sequence constrains the choice among what are assumed to be "synonymous" codons in immunoglobulin genes. Mutational hotspots in the variable region encode serine using AGY (Y indicates pyrimidine, i.e., C or T), while serine is encoded by TCN (N is any nucleotide) in the constant region.

The choice among synonymous codons also is constrained when a DNA sequence must, in effect, carry a label. As described in chapters 6 and 7, some bacteria

"mark" their DNA. For example, in certain bacterial genomes codon choice is constrained by overrepresentation of sequences that facilitate uptake of DNA from closely related bacteria. While only eight copies of a nine-base uptake signal sequence would be expected by chance, there are 1500 copies of that sequence in the *H. influenzae* genome.

Similarly, chapter 15 describes the constraints on codon choice due to exon splice enhancers and information at the exon–intron boundary, which mark the sequence as an exon and define its splice site.

Structural Information

We talk about DNA sequences, and our computer algorithms search them, as if nucleotides were letters that can be read much like those printed on this page. But after reading chapter 1, no reader is likely to view a DNA sequence simply as a one-dimensional string of "letters" again. The distinct tilt and twist of different steps of stacked base pairs along the helix may tend to average out, but for sequences highly enriched in one base, or in which a sequence motif is repeated, there can be substantial deviations from the average coordinates that are used in text-book models of the double helix. The biological consequences of such sequence-dependent variation in DNA structure are discussed in several chapters of this book.

For example, as described in chapter 13, sequences that are as high as 70–80% AT are highly "bendable" and are thought to affect chromatin structure and genome rearrangements in ciliates. As described in chapters 4 and 12, repeats of sequences may favor or inhibit nucleosome assembly, with effects both on gene expression and on the likelihood of genetic exchange between two parental genotypes at that site during meiosis.

As described in chapter 1, protein binding to DNA can be affected by effects of DNA sequence on backbone geometry. In fact, changing the sequence of a four-base "spacer" between two binding sites can change protein affinity by over three orders of magnitude. Similarly, the nature of the spacer sequence affects the efficiency of recognition of immunoglobulin recombination signal sequences, as referenced in chapter 9.

As the number of bases that separate two DNA sequences changes, the two sequences move relative to each other around the helix axis. As shown in chapter 3, changing the relative three-dimensional orientation of the left and right sides of the promoter by changing the distance between them by just 1 or 2 bases can have dramatic effects on interaction with transcription factors and thus on gene expression.

The context of a DNA sequence affects its mutability, which depends not only on the presence and composition of repeats and the direction of replication but also on the nature of flanking sequences. For example, if the sequence between repeats is a palindrome, deletion can be more likely.

Chapter 11 describes unique four-chain structures formed when G-rich sequences separate from the double helix during replication or transcription of the opposite strand. Such G-rich structures can stall replication and increase recombination and cause lethal genome instability, but they have been captured by the ever-creative

vertebrate immune system to focus the region-specific recombination that is required for a regulated switch between immunoglobulin heavy-chain classes. Thus, within immunoglobulin heavy-chain introns, G-rich sequences imply a region of recombination that is regulated by transcription.

Different nucleotide sequences may share certain physical chemical characteristics; for example, both A·T and G·C base pairs, but not an A·G base pair, fit the steric requirements of the replicative DNA polymerase. Thus it was possible to be blinded by viewing DNA as "letters," until a palindromic pattern of hydrogen bond donors and acceptors in the major groove was identified as the formerly elusive consensus sequence for P-element insertion in *Drosophila*.[10]

Chapter 1 emphasizes that analyzing the physical chemical properties of DNA sequences is not as straightforward as looking up codons in a table. The natural curvature of runs of As is reinforced when A runs are on the same side of the helix and "cancels out" when they are on opposite sides. Chapter 1 points out, while describing the induction of positive base inclinations at the 3′ end of AAAAAA tracts repeated in phase with the helical screw of DNA, that structural deviation from the coordinates of the standard B-form DNA can be propagated into neighboring sequences; due to the importance of sequence context, the local shape and flexibility of a DNA sequence cannot be predicted simply by adding up the tilt and twist parameters of individual steps. This effect of sequence context can make the identification of nucleotide "consensus" sequences challenging.

Relationships

In examples ranging from mutable repeats to RNA editing, information often is represented in the relationship between sequences, rather than in the specific sequences themselves. Certain RNA strands can form alternative loop structures that can regulate transcription and translation in response to the presence or absence of other molecules in the cell. A classic example is the attenuation loop involved in regulation of the biosynthesis of tryptophan.[11] More recent important examples involve sequences in mRNA at which metabolite binding affects the choice between alternative RNA structures.[12]

The essential, extensive, and evolving biological roles of RNA transcripts that do not encode amino acid sequences have recently become breathtakingly apparent.[13] For example, as discussed in chapters 8, 13, and 14, complementary RNA sequences trigger sequence-specific gene and/or transcript silencing.

Repeats in DNA can be loci of genetic variation, as described in chapters 3, 4, 6, and 7, but certain inverted repeats are sites of stability. During DNA replication on the lagging strand, the tendency of nearly-palindromic sequences to form base-paired hairpins can lead to a high frequency of a specific subset of mutations that result from a tendency to "repair" inexact matches across the hairpin.[14] "Correction" by internal palindromes has been proposed to protect Y chromosome sequences.[15] Relationships between DNA sequences can enable them to interact in ways other than the standard double helix, as discussed in chapter 11. The importance of sequence relationships in RNA in forming biologically-important structures is well appreciated.[16]

Evolutionary Information: Probable Future Genomes

In most discussions of evolution, mutation is described as a random and generally harmful process, with the genome as its hapless victim. However, during evolution, one genome sequence does become other, well-adapted, genome sequences, as a result of multiple mutations.

When genes and mutation were incorporated into evolutionary theory during the first half of the twentieth century (it should be noted that this was not only prior to genome sequencing but also prior to understanding how DNA encodes information or even that DNA is the genetic material), it was assumed that mutation was "random." Of course, those were not Darwin's words, as the concept of genes and the biochemistry of mutations were unknown when he proposed that biological evolution results from selection acting on variation in traits that are inherited.

As described in this book, nucleic acid sequences often change in ways that are not completely random. Intrinsic sequence-dependent variations in DNA sequence context (chapter 1) and in the structure and fidelity of enzymes responsible for replication and repair of DNA molecules (chapters 2, 3, and 6) result in sequence-dependent variations in the types and rate of mutations.

While it is clear that mutation is not random with respect to DNA sequence context, it has been assumed that mutation must be completely random with respect to its potential effects on phenotype because "selection lacks foresight."[17] However, because the world is not completely random, selection can gain a degree of "foresight"—to the extent that a lineage repeatedly must survive the same classes of challenges, such as pathogens surviving attacks by our immune response and our surviving attacks by pathogens.

Indeed, despite the expansive landscape of possible random changes, genomes often find repeated paths to the same solution. Chapter 8 presents a case study of adaptation of yeast to glucose restriction in which multiple clones shared the same breakpoint, at a transposable element. (This is one of many examples of how transposons can be valuable to the genome that hosts them.) Repeated (and reversible) amplification of certain sets of genes can enable stressed bacteria to reach beyond the limit of maximal expression of a single copy of these genes.

Most research in biology focuses on mechanisms that enable an organism to survive through one life cycle, including adaptation to changes in the environment that occur within its lifetime (such as the appearance of lactose). Yet genomes survive through an unbroken chain of living beings across evolutionary timescales. Lineages must survive challenges that may be repeated (such as climate cycles) or extended (such as the ongoing evolution of other organisms in the community). Some, but not all, chapter authors and I suggest that the evolutionary success of certain lineages may result from their genome being more efficient than random at exploring variation that is aligned with the nature of those challenges and opportunities their lineage has faced repeatedly during evolution, which underlies a phenotype of robustness to those classes of challenges[18]; this concept is illustrated by the rapid variation of pathogens' surface antigens (chapters 3 and 5).

Intriguing observations in the literature suggest that selected paths of exploration might exist for systems as diverse as toxin genes for cone snails[19] and, as discussed

in chapter 4, skeletal structure in dogs. Such observations should inspire experimental investigation of possible facilitated genomic exploration of lineage-essential traits, such as (pure but irresistible speculation) beaks upon which birds depend for access to available seeds.

As described in chapter 4, natural selection will act indirectly on the mechanisms that generate genome variation, much as it acts directly on beaks and wings. Darwin argued: "Why … should nature fail in selecting [useful] variations …? I can see no limit to this power." [20] This power can include selection for (useful) variations of distinct types of mutation along a strand of DNA.

Selection for Mechanisms That Generate Diverse Descendants

The best measure of "fitness" often is the ability to generate diverse descendants. As described in chapter 3, from the perspective of pathogens, hosts are an unsettled landscape. Which bacterium is the "fittest" can be redefined quite suddenly, such as by the appearance of a new antibody.

Reversible mutations generate diversity that enables a lineage to survive when one trait causes a disadvantage, without losing from the lineage a trait that may prove advantageous to descendants under other likely circumstances (chapters 3 and 6). For example, a capsule that shields bacteria from attack by the host's complement system prevents the bacteria from adhering to certain tissues. If the capsule appears and disappears through reversible mutations in repetitive sequences, bacteria will have descendants with and without the capsule, starting from either phenotype. In other words, bacteria resistant to complement will have a significant percentage of progeny that can stick to host tissues, while bacteria that can stick to host tissues will have a significant percentage of progeny that are resistant to complement. Monoallelic expression is another mechanisms that "hides" information (one allele in a diploid organism) that will predictably reappear in a subsequent generation (chapter 16).

As reviewed in chapter 12, two well-known mechanisms that ensure diversity in eukaryotic offspring are independent assortment of maternal and paternal contributions to the genome and recombination between homologous parental and maternal chromosomes during meiosis (with unexpected ratios of offspring phenotypes resulting from unrepaired heteroduplexes).

Frameworks for Exploration

An evolved infrastructure facilitates DNA mobility and recombination at sites where it is more likely to facilitate combination of functional patches of DNA, whether they are genes, regulatory regions, genomic contexts, or exons, rather than damage the recombining sequences.

Frameworks for Gene Movement in Prokaryotes

Chapter 7 describes the "metagenome," comprising potentially valuable information that flows horizontally among bacteria through a well-evolved infrastructure.

Transposable elements, phage, conjugative plasmids, and transposons all participate in the transfer of information among bacteria. Integrons, which carry a promoter and can capture and release gene cassettes, provide an efficient framework for transferring and expressing information; such a set of cassettes traveling together on one integron may carry resistance to all classes of clinically useful antibiotics, as well as to antiseptics and disinfectants. The spread of integrons can be aided by their incorporation into plasmids that have a broad host range.

As described in chapter 7, "genomic islands" combine relatively conserved core regions with other regions that contain a more variable set of genes that is appropriate to the environment of the particular strain or species.

Recombination and Movement of Information in Eukaryote Evolution

As described in chapter 9, the vertebrate immune system, which appears to have emerged through the creative action of a transposable element in the germline, has evolved a genomic infrastructure of gene segments, flanked by specific signal sequences, which are recombined according to specific rules. This genomic framework generates diversity that is aligned with the requirements of the protein; specifically, variation tends to occur where it affects the potential to bind to diverse pathogens without impinging upon effector functions required for pathogen disposal. While the repertoire of antibodies, with distinct pathogen-binding specificities that are many orders of magnitude larger than the currently annotated number of protein-coding genes in the human genome, evolves within the lifetime of each individual, it does so within a framework that was selected over evolutionary time.

Based on the clear evidence of such biochemistry being available to the immunoglobulin genes, it is reasonable to consider that information that aligns the probability of distinct types of variation along a nucleotide sequence with the requirements for protein function might be a characteristic of other "successful" gene families too, facilitating expansion to large numbers of functional members by not depending exclusively on chance and random variation to sculpt each new duplicated copy of the gene.[21] In fact, chapter 9 suggests that hypermutation of the Ig variable regions may have predated in evolution the appearance of their extraordinary somatic recombination.

Chapter 11 suggests that G-rich sequences, which define the sites of exon switching in the immunoglobulin genes, might stabilize four-stranded DNA structures in the germline, which could, through ectopic recombination, facilitate exon shuffling (the idea, first proposed when introns were discovered, that functional domains might move around the genome, thus facilitating an exploration of potential new combinations of information). That G-rich sequences do recombine in the germline is suggested by yeast, in which a meiosis-specific protein that binds G-rich four-chain structures promotes the interaction of two helices. In fact (as described in chapter 12), certain GC-rich sequences create regions distributed over approximately 100 to 500 base pairs that pairs that are hotspots of recombination in meiosis.

Chapter 9 presents several lines of evidence for ongoing activity in germ cells of the transposon-derived enzyme RAG (recombination activating gene), which is involved in immunoglobulin gene segment rearrangement. The continued role of transposable elements in eukaryote genome evolution, including two mechanisms for exon shuffling, is discussed in chapter 8. As transcription of elements that transpose through an RNA intermediate often continues into flanking host DNA, such elements can contribute to exon shuffling due to their propensity to carry host sequences; these transcripts return to the nucleus and can prime reverse transcription into chromosomal DNA at sites nicked by a transposon-encoded endonuclease. Thus a transcribed element potentially can carry, to another place in the genome, any sequence that is its 3′ neighbor. Such exploration of new combinations of functional pieces of DNA can lead, for example, to adding a regulatory domain to an enzyme, which might be one step in linking together a control network, as described in chapter 8.

Even without "jumping," transposable elements can facilitate exon shuffling, by providing homology that enables ectopic recombination between otherwise unrelated sequences in their neighborhood, much as recombination at G-rich sequences described in chapter 11 does not require extensive sequence homology. Whether transposons move around a genome through an RNA intermediate, directly as DNA, or are no longer "jumping," they spend most of their time as DNA, part of the genome. While the probability of ectopic recombination between any two elements may be low, in aggregate, the one million *Alu* elements in the human genome are likely to contribute significantly to genome evolution. Chapter 8 reports that *Alu* elements, dispersed through the genome, also can become new exons.

In addition, both trypanosomes (chapter 5) and chicken immunoglobulins (chapters 9 and 10) point to the important role of pseudogene fragments, which contribute diverse sequences by gene conversion. Pseudogenes are widespread; there is growing evidence for the contribution of information derived from pseudogenes both to gene evolution and to regulation of gene expression.[22]

Genome Organization: The Importance of Neighborhood and Context

Both the mutability and the "meaning" of a nucleotide sequence depends on its context. For example, a T is more likely to be deleted or added in a run of other Ts. And as for that nucleotide run, whether TTTTTTTTT "means" phenylalanine–phenylalanine–phenylalanine or a specific level of gene expression depends upon whether the mononucleotide run is in a "protein coding" region or separates the left and right side of a bacterial promoter. The effect of context can be felt over a range of distances, from neighboring twists of the helix (chapter 1) to chromosomal regions (chapters 8 and 16).

Gene Expression Is Affected by Neighboring Transposons

With the potential to involve that gene in a regulatory hierarchy through RNAi-mediated generation of heterochromatin, Chapter 8 describes two routes by which

transposon-derived sequences in the neighborhood of a gene can lead to dsRNA. Transcription initiated in a host sequence can generate dsRNA by reading through an element that contains terminal inverted repeats. Alternatively, transcription initiated in a transposable element can generate dsRNA if its promoter drives transcription into a nearby gene on the antisense strand. (When the neighboring gene is transcribed, sense and antisense transcripts anneal, forming dsRNA.) Thus genes that neighbor the same transposable element yet are scattered through the genome might be coordinately silenced by an environmental trigger that affects the element's promoter.

Regions of Monoallelic Expression

We are used to thinking that as humans we are strictly diploid, with two active copies of every gene (except those on the X and Y chromosomes)—one inherited from our father and one from our mother, but which now are equivalent and both equally "ours." As described in Chapter 16 however, one copy of a parentally imprinted locus is turned off, depending upon the parent of origin. We are, in effect, haploid at that locus, with the information on the silent copy just passing through us to be revealed in a future generation. Imprinting occurs with distinct patterns, depending upon whether the genome is developing in the male or female germline. Imprinted genes "remember" whether they are from Dad or from Mom. An imprint may be removed in some but not all tissues in the progeny, most notably the developing germline, where it then is re-set according to the sex of the future new parent. The "reasons" for this are the subject of active speculation, but, as discussed in chapter 16, parental imprinting may have its biochemical origin in the requirement of homologous chromosomes to carry parental marks to enable appropriate sorting during meiosis and mitosis. Because whether a gene is imprinted depends upon the genomic context in which it is placed, chapter 16 leads into a literature that discusses the effects on gene activity of information that defines chromosome neighborhoods.

As described in chapter 16, many additional loci in diploid genomes that experience monoallelic expression (also termed "allelic exclusion"), including the T-cell receptor, natural killer receptors, and receptors for odors and pheromones, encode proteins that interact with agents from the environment. An "exposed" suddenly haploid allele will experience direct selection. For example, as discussed in chapter 10, antibody-producing cells undergo selection based upon the binding activity of the expressed allele. As discussed in chapter 5, only one specialized expression site for trypanosome surface antigen genes is transcribed at a time; an active site becomes silenced and an inactive site activated through an as yet mechanistically undefined interaction between the two.

Regulated, Region-specific Recombination

While much recombination is site-specific, some recombination is increased along a region of DNA. In some cases it is clear that such "region-specific recombination" is regulated through transcription. In particular, as described in chapter 11,

introns that participate in the immunoglobulin class switch are enriched in G-rich regions that attract recombination; recombination occurs in response to extracellular signals that stimulate transcription across the G-rich region.

Chapter 12 describes the requirement of specific transcription factors, but in contrast to the immunoglobulin class switch no evidence to date for transcription itself, at hotspots of meiotic recombination. Thus, many proteins that have been considered "transcription factors" may have as their primary role the regulated opening of chromatin structure, whether for transcription or for other purposes related to genome organization and handling.

As discussed in chapter 5, the probability of recombination is affected by position on a chromosome as sequences in subtelomeric regions have an increased probability of ectopic recombination. While some genomic regions are hotspots of recombination and/or foci of genomic flux, there also can be great local stability. As described in chapter 12, genome sequencing has led to investigations into the nature of recombination hotspots and of cold spots that may enable certain blocks of genes to remain together from generation to generation. Certain gene regions, such as the HOX genes, experience duplications as a block, but maintain their linear relationship to each other within these blocks across long evolutionary timescales.[23] In contrast, there is a particularly high rate of germline recombination at specific loci in the human genome, such as at the HLA locus,[24] at which diversity is especially important.

The precise location of meiotic recombination within a meiotic recombination hotspot depends upon the specificity of Spo11, the endonuclease that makes the initiating double-strand break. The location of sites favored by this and other nucleases, such as the endonuclease involved in priming and inserting sequences following reverse transcription (chapter 8), also can fall under selection.

Contexts Defined by the Timing of Replication

Preliminary work links the boundaries between "early" and "late" replicating regions with the boundaries of regions with a tendency to amplify during tumor formation and boundaries of syntenic regions that have remained together across evolutionary time,[25,26] indicating that these boundaries have biological, or at least biochemical, identities.

In a wide range of species, whether a region replicates "early" or "late" appears to be one way of defining sequence regions with distinct properties. Chapter 16 indicates that the timing of replication at imprinted loci differs depending on the parent of origin. Chapter 13 points out that regions of DNA that are destined to be left out of the somatic nucleus of ciliates replicate during a second phase of DNA synthesis. Chapter 12 observes that many cold spots of meiotic recombination replicate late during pre-meiotic replication when chromatin alterations occur that subsequently influence region-specific double-strand break formation during meiosis. Chapter 7 observes that sequence conservation is much higher near the origin of replication than at the terminus in prokaryotes.

When distinct regions of the genome replicate at different times or in different territories, nucleotide pools and relative levels of mismatch repair proteins (chapter 6),

polymerases with distinct fidelity (chapter 2), and other protein activity, all can differ giving these regions their own spectrum of mutations (chapter 2). Combined with sequence-dependent variations in DNA structure (chapter 1), there can be distinct region- and time-dependent effects on the mutability of different classes of sequences, such as mononucleotide or tetranucleotide repeats (chapter 3). This provides a biochemical route by which multiple genes and regulatory pathways can experience selection through effects on the mutation phenotype.[27]

To the extent selective pressure has operated on variations in the probability of mutation along a DNA sequence, not only will classes of sites in a genome have different probabilities of distinct types of mutation, but those classes that have a higher probability of success will tend to become more frequent than those that are more consistently deleterious.

Regulated Variation in Nucleic Acid Sequences

Much like the regulated alterations in nucleic acid sequence involved in alternative splicing (chapter 15) and ciliate genome reorganization (chapter 13), certain classes of mutations in nucleotide sequences are affected by the levels of distinct gene products and therefore can be regulated.

Effects of Gene Products on Mutation: Lessons From the V-region

As described in some detail in chapter 10, changes in the balance of activity of specific proteins can change the outcome of a mutagenic insult. For example, C to U deamination occurs spontaneously at a significant rate, estimated to be 100 times per cell per day in the human genome.[28] The immune system has captured this physical chemical property of the nucleotide C, accelerating it enzymatically at sites in the variable region. While mechanisms that repair deaminated C evolved long ago, and are present in bacteria, enzymatic C deamination in the immunoglobulin variable region results in mutations; regulation of repair enzyme activities can affect the nature of the genetic change that results. Since deamination occurs spontaneously throughout the genome, mutation could be affected at any site in the genome not only by targeted enzymatic deamination but also simply by manipulation of repair.

Effects of Gene Products on Variation at G-rich Sequences

As described in chapter 11, an extracellular signal can induce targeted genome rearrangement in B cells by inducing RNA transcripts that initiate from promoters located within immunoglobulin heavy-chain introns. These transcripts do not encode proteins but rather free a G-rich region from the double helix, initiating recombination between two transcribed regions. Thus, at least theoretically, regulated changes in the activity of proteins that interact with G-rich DNA could change the level of genome instability in the germline as well. Indeed, as described in chapter 11,

decreased levels of helicases that unwind four-strand structures formed by G-rich DNA do lead to increased genome instability; inhibition of another helicase-like protein leads to genome-wide destabilization of polyguanine tracts in nematodes. Similarly, deficiency of a protein that, at its normal evolved level, prevents cotranscriptional hybrid formation in yeast leads to hyperrecombination that is associated with actively transcribed regions of DNA.

Effects of Temperature and Nutrient Levels

In yeast, as described in chapter 12, the locations of hotspots of meiotic recombination are altered in different environments due to yeast hotspot dependence on transcription factors (and thus on the metabolic states that regulate them). The observation that temperature determines the rate of site-specific recombination in the locus encoding the *E. coli* fimbrial proteins, as described in chapter 3, indicates that environmental signals may not only inform a bacterium that it has arrived in its host but also, as a response, trigger it to access specific genomic sequences that are available in its implied repertoire ("species genome") and which are appropriate to the host environment.

Genomic Change Mediated by Mobile Elements

Barbara McClintock suggested[29] that stress could activate genome reorganization, which we now might describe in molecular terms by saying, for example, that a transposase that recognizes a class of genomic sequences might be induced by a biochemical signal of stress. Perhaps this explains why during adaptation of rice under the stress of selection by humans for growth in temperate climates there was a rapid increase observed in the number of copies of a transposable element that inserts into regions where it may alter the regulation of rice genes, as described in chapter 8. Similarly, chapter 4 points to a repeated contraction of a $(CT)_n$ repeat that is reported to occur only in wheat exposed to the head blight pathogen.

In bacteria, stress activates the SOS response. Prophage, which can facilitate the movement of genes among bacteria, as described in chapter 7, are induced by stress. As described in chapter 3, slippage rates of repeat tracts, such as those found in the genes encoding *N. meningitidis* lipopolysaccharide biosynthetic enzymes, depend on the activity of distinct proteins and so can be regulated. The rates are affected by transcription, the SOS response, and by environmental signals, as well as by mutations with sequence-context effects, such as those that perturb leading and lagging strand DNA synthesis.

As described in chapter 6, a population-level bacterial survival skill results from the presence, in the sequence of mismatch repair genes, of closely spaced repeats that tend to lose and regain mismatch repair activity through recombination. As described in chapters 3, 6, and 12, the diversity of sequences that can be recombined into a genome is increased when mismatch repair is decreased. Chapters 6 and 7 describe ways in which bacterial competence for, and acceptance of, DNA present in the environment itself can be regulated by environmental signals.

A Genome Is Not a Bag of Letters

The number of genome sequences available for analysis suddenly has become close to overwhelming. In addition, many new high-throughput laboratory techniques will inform issues raised in this volume. For example, as chapter 12 points out, the availability of whole-genome microarrays will overcome the challenge of obtaining statistically significant data regarding the probability of recombination across a genome, including at cold spots.

However, as we ramp up our data gathering and crunching, it is important to note that genomes are not just bags of genes, and genes are not just strings of context-independent letters. A nucleotide's "meaning" and behavior depends on its relationship to neighboring sequences and on the nature and relative levels of proteins expressed under distinct circumstances in that genome. Therefore, we must be cautious in drawing broad conclusions about adaptation from experimental studies that use laboratory constructs rather than genes with their own mix of "synonymous" codons and in the context in which they evolved. As discussed in chapter 6, bacteria that lose mismatch repair activity through recombination between closely spaced repeats may regain activity by recombination at these sites with exogenous DNA. Thus they are able to use the mutator phenotype as a temporary step to adaptation without then having to evolve new mismatch activity from scratch. Simply deleting a mismatch repair gene in the laboratory does not duplicate the endogenous behavior.

Similarly, as pointed out in chapter 5, laboratory-adapted strains of trypanosomes switch surface antigens at rates that are far lower than those observed during infection; and, when grown in the laboratory, *B. burgdorferi* tends to lose the plasmids required for generation of diversity through gene conversion.

Added to these observations is our growing appreciation for the regulatory role of RNA transcripts that do not encode protein, which serves to emphasize the importance in biological regulation and evolution of complementary relationships between sequences. Therefore, in spite of the great success of experiments that have cut and pasted DNA, it is now time to consider, in experimental design, what we might learn by studying sequences in the contexts in which they evolved.

Summary: A Sequence That Implies Other Sequences

For over half a century, we have been in the thrall of the double-helical structure of DNA, which, in an instant, revealed that information can be transferred between generations by a simple rule: A pairs with T, G pairs with C. In its beautiful simplicity, this structure, along with the table of codons worked out in the following decade, had entranced us into believing that we can fully understand the information content of a DNA sequence simply by treating it as text that is read in a linear fashion. While we have learned much based on this assumption, there is much that we have missed.

Now, at the beginning of the twenty-first century, with entire genome sequences appearing before us, biologists can appreciate the feeling experienced by physicists at the beginning of the twentieth century, of the firm ground slipping out from under their feet. We are indeed in a "new world."

Vast unannotated spaces appear in genome sequences, then fill up with transcripts that do not encode proteins. Far from being a passive tape running through a reader, genomes contain information that appears in new forms, and which creates regions with distinct behavior. Some genome regions are "gene rich," some mobile, some full of repeats and duplications, some sticking together across long evolutionary distances, some readily breaking apart in tumor cells. Even protein-coding sequences can carry additional information, taking advantage of the flexible coding options provided by the degeneracy of the genetic code. When viewed at the level of the RNA transcript or the DNA itself, "synonymous" codons are not always synonymous. Even something as familiar as an RNA transcript with alternative splice forms bears information "about" itself: defining exons and introns, and framing out how it will be spliced under different circumstances. There are new concepts to capture from the flood of often surprising new data.

Even the most senior among us has become a student again, working to learn from this rich font of new information; unless we remain open to the possibilities of discovering new types of information, and novel ways of encoding information, we may fail to perceive such information when it appears before us.

The chapters in this volume touch on one or more of three interconnected themes: information can be implied, rather than explicit, in a genome; information can lead to focused and/or regulated changes in nucleotide sequences; information that affects the probability of distinct classes of mutation can have implications for evolutionary theory.

Rather than simply summarize chapter by chapter in this overview, I have worked to integrate these observations into a conceptual framework. Whether or not a reader (or a chapter author!) is intrigued with the extent to which variations in the probability of mutations enable the ability to evolve itself to evolve under selective pressure, what is clear is that there are "deeper" ways to look at and understand nucleotide sequences and sequence contexts than had been obvious from viewing base sequences solely as linear strings of letters.

All of us who have contributed to this volume anticipate that this book will inspire readers to ask their own challenging questions, and to suggest their own syntheses of the unexpected, unexplored, and unexplained information in the rapidly expanding genomic databases. New terms are suggested, such as "implied information," "metagenome," and "species genome." At times even such familiar words as "information" and "code" seem to constrain our ability to describe what is represented in the genome. The Epilogue proposes adapting the concept of "protocol," which is widely used in describing another dynamic, robust, and evolving entity: the internet. Even our view of what can be "inherited" is challenged, as the population descendant from an individual bacterium inherits diversity from one specific manifestation of diversity at each highly-mutable locus, enabling survival in a broader range of niches than the individual parent itself could survive.

As this volume ranges across a broad field of biological problems it cannot be comprehensive, rather is intended provide a unique perspective on, and a doorway into, the literature while inspiring the reader's imagination to expand its appreciate of the types of information that may be represented in a genome. As you journey through the pages of this volume, may your reflections lead you to question

assumptions (including those of this editor) ask more challenging questions in your own research, and send you on a path to additional startling discoveries.

Acknowledgments I would like to thank the many readers—some known to the chapter authors, some anonymous, some chapter authors themselves—who contributed their time to review one or more chapters. I also would like to thank those who contributed advice on the selection of chapter authors, and those at Oxford University Press who contributed their hard work "behind the scenes" without whom this book would not be in your hands. Most of all, I would like to thank the chapter authors for their thought-provoking contributions to this volume.

Much of the time devoted to this book was exchanged for time I might have spent with family, including Rockella, Parker, Michael, and Brooks, and so this chapter is dedicated to them.

1

Sequence-Dependent Properties of DNA and Their Role in Function

Donald M. Crothers

Overview

The genome is a complex nuclear organelle whose function is to encode the information needed to maintain the living state as cells grow and divide, and as generations pass from parents to progeny. Much attention has focused on the DNA sequences that encode proteins, the workhorse molecules of biochemical metabolism and biological structures. However, these sequences account for only a fraction of the total human genome. Moreover, because of the degeneracy of the genetic code, there are many ways to encode a specific protein. Local DNA structure depends on sequence, as do the mechanical properties, such as bending and twisting flexibility. As a consequence, much more information is encoded in the genome than is accounted for by protein sequences. Deciphering the complex secondary code[s] has only begun.

Introduction

The B-DNA helix structure proposed by Watson and Crick[1–3] has been the dominant icon of molecular biology for half a century. Only purine–pyrimidine base pairs, A with T and G with C, are tolerated in the confines of the regular helical structure, shown in figure 1.1a,[4] as refined from fiber diffraction studies. Inherent in the structure is the logic of its replication, since each strand can serve as the template for synthesis of its complement. Incorporation of deoxyribose in the alternating sugar–phosphate backbone confers polarity on the strands: each sugar has a phosphodiester linkage on its 5′ and 3′ oxygens. By convention, DNA sequences are written in the direction that corresponds to the order of biosynthesis—from 5′ end to 3′ end. A key feature of the structure is that the two strands in the duplex are antiparallel: the 5′ end of one chain and the 3′ end of the other chain are at the same terminus of the duplex. The simplicity of a uniform base-paired helix, with 10 bp per turn, 3.4 Å rise per base pair, was a key element in the rapid acceptance of the structure, and the explosive growth of molecular biology that followed.

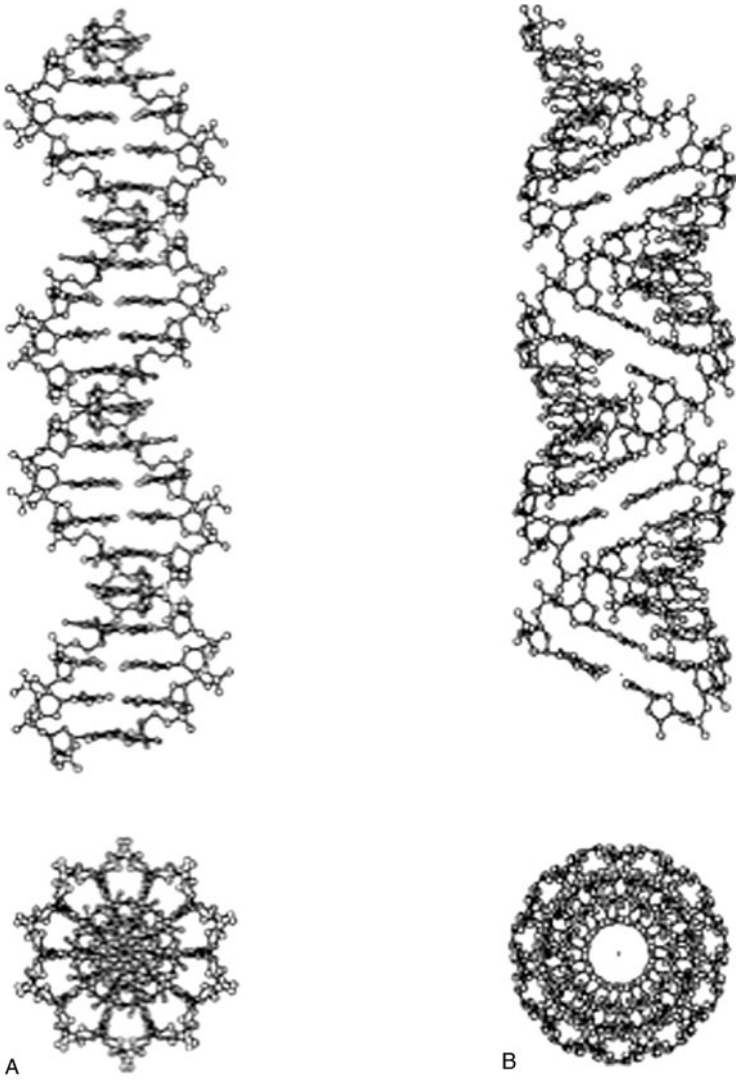

Figure 1.1. DNA structure. Mutually perpendicular views of (A) the "average" calf thymus B-DNA structure, determined from fiber diffraction at high humidity, and (B) A-DNA, determined at low humidity. (Reprinted from reference 4.)

However, 50 years later we recognize that the B-DNA helix shows significant variations in properties, depending on the base sequence. A thermodynamic property that became evident early on[5] is the higher stability conferred against melting or denaturation by a high content of G·C base pairs, readily explained by the greater resistance to breaking base pairs provided by the three hydrogen bonds in a G·C pair compared with the two bonds in an A·T pair (figure 1.2).[6] As a consequence, a DNA molecule that contains 100% G·C pairs has a melting temperature,

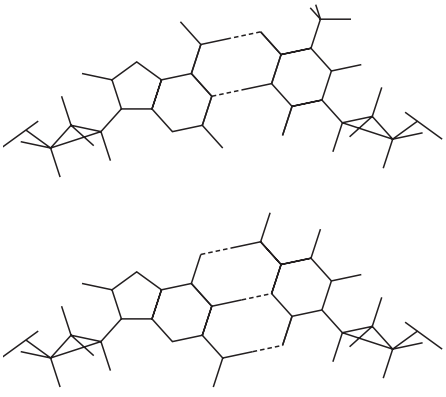

Figure 1.2. Standard base pairing geometries. (Reprinted from reference 6.)

or T_m, about 40 °C higher than a molecule containing only A·T pairs (when both are longer than a few hundred base pairs). X-ray and NMR structural studies over many years have revealed substantial variations in local structure from one duplex to another. Further elements of complexity and variety of structure are contributed by superhelical conformations of circular DNAs,[7] dramatic global structural alterations that yield three-stranded[8] and four-stranded structures[9] (see also chapter 11). Formation of these multistranded structures depends on sequence correlations such as runs of pyrimidines on one strand and purines on the other, or on repeated tracts of G residues along one strand in telomeric sequences. Branched nucleic acids also provide important intermediates in recombination and replication.[10]

Another key feature to keep in mind is that only about 5% of the human genome is transcribed into mRNA. Small RNAs with a variety of functions account for an additional percentage. Does the rest of the genome encode essential information? The answer is clearly yes for regulatory regions, where proteins bind sequence-specifically. While the hydrogen bonding functionalities of the DNA bases are clearly important for binding specificity, variations in local DNA structure and in mechanical properties such as bending and twisting stiffness also play important roles. Packaging of DNA from nucleosomes through chromatin and up to mitotic chromosomes also depends in important ways on local structure and mechanical properties. Hence the information content of human DNA is contained not only in the base pair complementarity rules but also in the more subtle sequence-dependent structural and mechanical properties. Deciphering this code is still in its early stages.

Local Structural Variations

How DNA Structure Is Characterized

At a meeting at Cambridge University in 1988, a group of nucleic acid structural biologists agreed on a standard set of parameters to characterize DNA at a local structural level. These are shown diagrammatically in figure 1.3.[6] The top set of

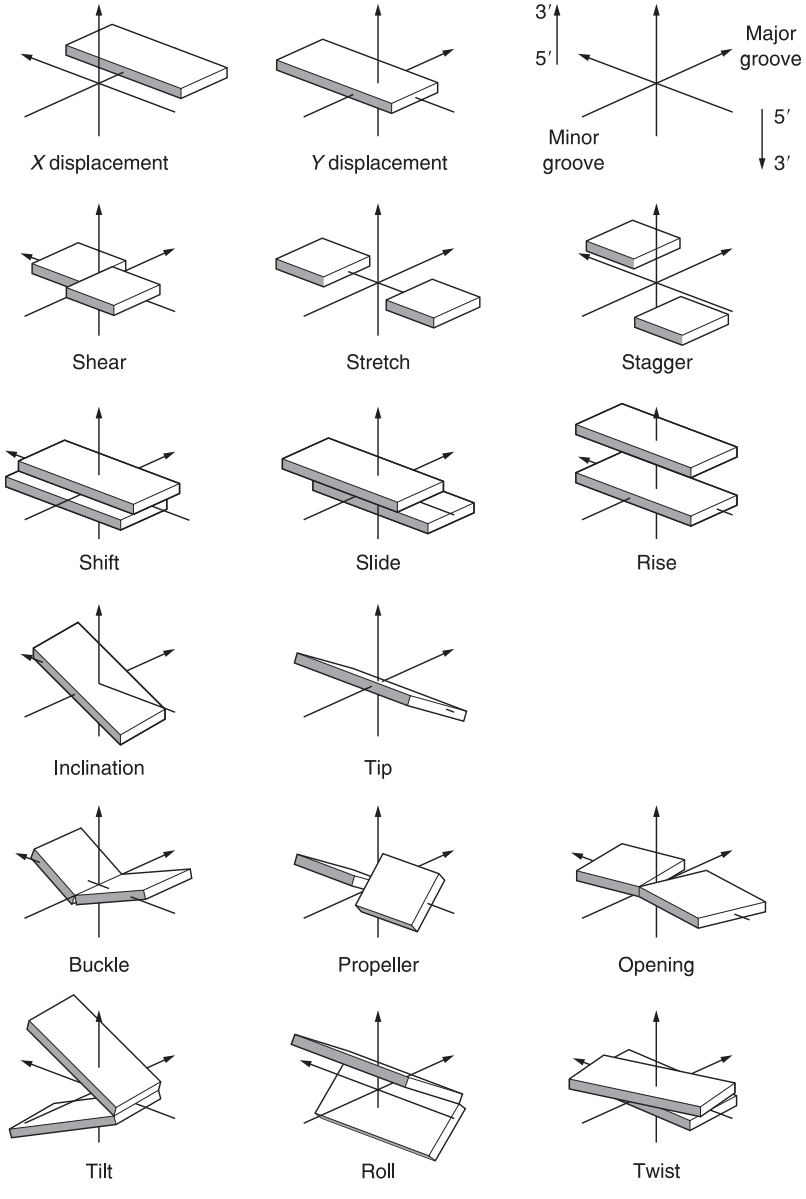

Figure 1.3. Helical parameters. Translations are shown in the upper part of the diagram and rotations in the lower part. Each section contains base-pair axis, intra-base-pair, and inter-base-pair parameters. (Reprinted from reference 6.)

eight parameters represent distances; they are organized into base-pair movements (top row) relative to the axis, intra-base-pair distances (middle row) and inter-base-pair distances (bottom row). The bottom set of eight parameters represent angles, similarly organized into rows. For reference, most of these quantities are

near zero for the classical B-DNA structure, excepting rise (3.38 Å) and twist (36°). More recent refinements of fiber diffraction data[6] for B-DNA have nonnegligible values of propeller twist, −13° to −15°. The classical A-form structure (figure 1.1b), on the other hand, has substantial positive roll (6.3°), positive inclination of the base pairs (12°), substantial *x*-displacement (4.1 Å), which moves the base pairs away from the helix axis, and slide between base pairs of 1.6 Å. A number of papers on nucleic acid structure that review this field in detail are provided in the *Oxford Handbook of Nucleic Acid Structure*.[11]

Sequence-dependent Variation of Structure in DNA Duplexes

Several approaches have been taken to determine experimentally the sequence-dependent variation of DNA structural parameters. The most direct method is x-ray crystallography of oligonucleotide duplexes, for which there is now a substantial database.[12,13] Gorin et al.[14] provide a useful summary of the major parameters for the 10 independent dinucleotide base-pair steps, and point out some significant correlations between them. (There are 10 independent dinucleotides instead of 4^2 (= 16) because of complementarity; for example, 5′–A–C–3′ is complementary to 5′–G–T–3′ in a duplex context, and hence one implies the other. Four of the dinucleotides, such as 5′–A–T–3′, are self-complementary, leaving 12 dinucleotides that provide six independent base-pair steps.) Averaged over the 10 dinucleotides, for a data set of 195 structures, roll and tilt are both nearly zero (–0.2° and 0.0°, respectively). However, there are significant variations from one dinucleotide to another. For example, the average roll for the G–C dinucleotide is −7.0°, whereas roll for G–G is +6.5°. Tilt generally varies less than roll,[15] by roughly a factor of two. This accords with simple mechanical expectation that it should be easier to rotate about the base pair long axis (roll) than about the short (tilt); see figure 1.3.

Twist is another parameter that varies considerably among dinucleotides, from a low of 30.5° for A–G to a high of 40.0 for T–A. However, it is generally not possible to build a severely underwound helix from these dinucleotide steps because the low or high value of the starting dinucleotide is generally compensated by a correspondingly high or low value of the next dinucleotide. For example, T–A–T has an average twist of 36.7°, much like the average value of 35.8° for A–A–A. Similarly, the average twist for A–G–A is 35.0°.

The structural parameters for DNA duplexes, such as those in the previous paragraph, should not be thought of as rigid values that can be applied to every stretch of DNA. In fact, these parameters vary depending upon on sequence context. In addition, structural differences can arise in the measurement itself, as a consequence of variable crystal packing forces. The root-mean-square (rms) variation, or square root of the variance, around the average value for a particular dinucleotide, averaged again over all dinucleotides, is 5.7° in the Gorin et al.[14] data set. The corresponding value for tilt is 3.6°, in agreement with the greater difficulty of varying tilt from its mechanical equilibrium value. These values accord quite well with the thermal fluctuations in the sum of roll and tilt deduced from the measured bending flexibility of DNA molecules in solution, as discussed below. The observed