Jerome K. Percus

Cambridge Studies in Mathematical Biology

Mathematics of Genome Analysis

MATHEMATICS OF GENOME ANALYSIS

JEROME K. PERCUS
New York University

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011-4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published 2002
Reprinted 2004

Printed in the United Kingdom at the University Press, Cambridge

Typeface Times Roman 10.25/13 pt. *System* LaTeX 2$_\varepsilon$ [TB]

A catalog record for this book is available from the British Library.

Library of Congress Cataloging in Publication Data
Percus, Jerome K. (Jerome Kenneth)
Mathematics of genome analysis / Jerome K. Percus.
p. cm. – (Cambridge studies in mathematical biology ; 17)
Includes bibliographical references and index.
ISBN 0-521-58517-1 – ISBN 0-521-58526-0 (pb.)
1. Genetics – Mathematical models. 2. Genetics – Statistical methods. 3. Gene mapping – Mathematical models. 4. Gene mapping – Statistical methods. I. Title. II. Series.
QH438.4.M3 P47 2001
572.8′6′0151 – dc21 2001035087

ISBN 0 521 58517 1 hardback
ISBN 0 521 58526 0 paperback

MATHEMATICS OF GENOME ANALYSIS

The massive research effort known as the Human Genome Project is an attempt to record the sequence of the three billion nucleotides that make up the human genome and to identify individual genes within this sequence. Although the basic effort is of course a biological one, the description and classification of sequences also naturally lend themselves to mathematical and statistical modeling.

This short textbook on the mathematics of genome analysis presents a brief description of several ways in which mathematics and statistics are being used in genome analysis and sequencing. It will be of interest not only to students but also to professional mathematicians curious about the subject.

Jerome K. Percus is Professor of Physics and Mathematics at the Courant Institute of Mathematical Sciences and Department of Physics at New York University, where he has taught since 1958. He has held visiting positions at Middlesex Hospital Medical School in London, Columbia University, Rutgers University, Princeton University, Rockefeller University, Yukawa Institute in Kyoto, Tokyo University, Norwegian Institute of Technology, Max Planck Institute in Tubingen, Catholic University in Rio de Janeiro, Ecole Polytechnique in Lausanne, Soviet Academy of Sciences in Moscow, Leningrad, Kiev, and Lvov, University of Paris, Nankai University, and Tsinghua University in China. He has received the Pregel (New York Academy of Science), Pattern Recognition Society, and Hildebrand (American Chemical Society) Chemical Physics awards.

Cambridge Studies in Mathematical Biology

1 Brian Charlesworth, *Evolution in age-structured populations* (2nd ed.)
2 Stephen Childress, *Mechanics of swimming and flying*
3 C. Cannings and E. A. Thompson, *Genealogical and genetic structure*
4 Frank C. Hoppensteadt, *Mathematical methods of population biology*
5 G. Dunn and B. S. Everitt, *An introduction to mathematical taxonomy*
6 Frank C. Hoppensteadt, *An introduction to the mathematics of neurons* (2nd ed.)
7 Jane Cronin, *Mathematical Aspects of Hodgkin-Huxley neural theory*
8 Henry C. Tuckwell, *Introduction to theoretical neurobiology*
Volume 1, *Linear cable theory and dendritic structures*
Volume 2, *Non-linear and stochastic theories*
9 N. MacDonald, *Biological delay systems*
10 Anthony G. Plakes and R. A. Miller, *Mathematical ecology of plant species competition*
11 Eric Renshaw, *Modelling biological populations in space and time*
12 Lee A. Segel, *Biological kinetics*
13 Hal L. Smith and Paul Waltman, *The Theory of the chemostat*
14 Brian Charlesworth, *Evolution in age-structured populations* (2nd ed.)
15 D. J. Daley and J. Gani, *Epidemic modelling: an introduction*
16 J. Mazumdar, *An introduction to mathematical physiology and biology* (2nd ed.)
17 Jerome K. Percus, *Mathematics of genome analysis*

Preface

"What is life?" is a perennial question of transcendent importance that can be addressed at a bewildering set of levels. A quantitative scientist, met with such a question, will tend to adopt a reductionist attitude and first seek the discernible units of the system under study. These are, to be sure, molecular, but it has become clear only in recent decades that the "founder" molecules share the primary structure of linear sequences – in accord with a temporal sequence of construction – subsequent to which chemical binding as well as excision can both embed the sequence meaningfully in real space and create a much larger population of molecular species. At the level of sequences, this characterization is, not surprisingly, an oversimplification because, overwhelmingly, the construction process of a life form proceeds via the linear sequences of DNA, then of RNA, then of protein, on the way to an explosion of types of molecular species. The founder population of *this* subsequence is certainly DNA, which is the principal focus of our study, but not – from an informational viewpoint – to the exclusion of the proteins that serve as ubiquitous enzymes, as well as messengers and structural elements; the fascinating story of RNA will be referred to only obliquely.

That the molecules we have to deal with are fundamentally describable as ordered linear sequences is a great blessing to the quantitatively attuned. Methods of statistical physics are particularly adept at treating such entities, and information science – the implied context of the bulk of our considerations – is also most comfortable with these objects. This hardly translates to triviality, as a moment's reflection on the structure of human language will make evident.

In the hyperactive field that "genomics" has become, the focus evolves very rapidly, and "traditional" may refer to activities two or three years old. I first presented the bulk of this material to a highly heterogeneous class in 1993, again with modification in 1996, and once more, further modified, in 1999. The aim was to set forth the mathematical framework in which the burgeoning activity takes place, and, although hardly impervious to the passage of time,

this approach imparts a certain amount of stability to an intrinsically unstable divergent structure. I do, of course, take advantage of this nominal stability, leaving it to the reader to consult the numerous technical journals, as well as (with due caution) the increasing flood of semitechnical articles that document the important emerging facets of the current genomics field.

It is a pleasure to acknowledge the help of Connie Engle and Daisy Calderon-Mojar in converting a largely illegible manuscript to, it is hoped, readable form, of Professor Ora Percus for insisting on reducing the non sequiturs with which the original manuscript abounded, and of numerous students who not only maintained intelligent faces, but also gently pointed out instances of confusion in the original lectures.

Contents

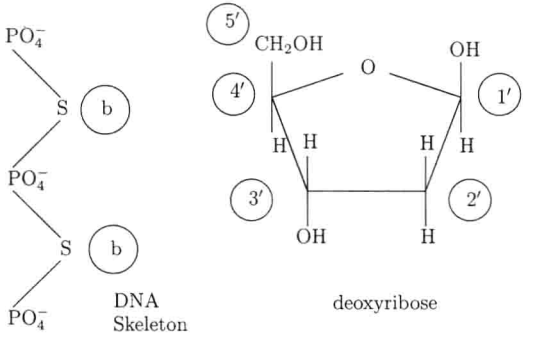
1

Decomposing DNA

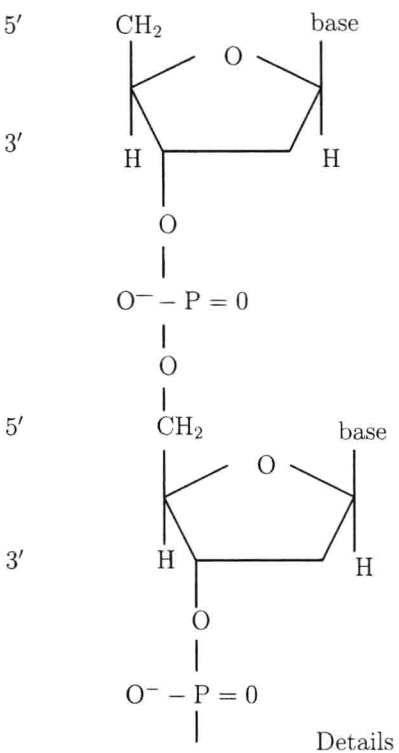
1.1. DNA Sequences

The realization that the genetic blueprint of a living organism is recorded in its DNA molecules developed over more than a century – slowly on the scale of the lifetime of the individual, but instantaneously on the scale of societal development. Divining the fashion in which this information is used by the organism is an enormous challenge that promises to dominate the life sciences for the foreseeable future. A crucial preliminary is, of course, that of actually compiling the sequence that defines the DNA of a given organism, and a fair amount of effort is devoted here to examples of how this has been and is being accomplished. We focus on nuclear DNA, ignoring the miniscule mitochondrial DNA.

To start, let us introduce the major actor in the current show of life, the DNA chain, a very long polymer with a high degree of commonality – 99.8%, to within rearrangement of sections – among members of a given species [see Alberts et al. (1989) for an encyclopedic account of the biology, Cooper (1992) for a brief version, Miura (1986), and Gindikin (1992) for brief mathematical overviews]. The backbone of the DNA polymer is an alternating chain of *phosphate* (PO_4) and sugar (S) groups. The sugar is *deoxyribose* (an unmarked vertex in its diagrammatic representation always

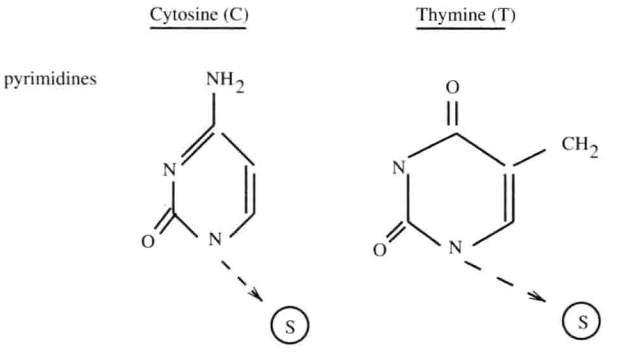


signifies a carbon atom) with standard identification of the five carbons as shown. Successive sugars are joined by a phosphate group (phosphoric acid, H_3PO_4, in which we can imagine that two hydrogens have combined with 3′ and 5′OHs groups of the sugar, with the elimination of water, whereas one hydrogen has disappeared to create a negative ion); the whole chain then has a characteristic 5′–3′ orientation (left to right in typical diagrams, corresponding to the direction of "reading," also upstream to downstream). However, the crucial components are the side chains or bases

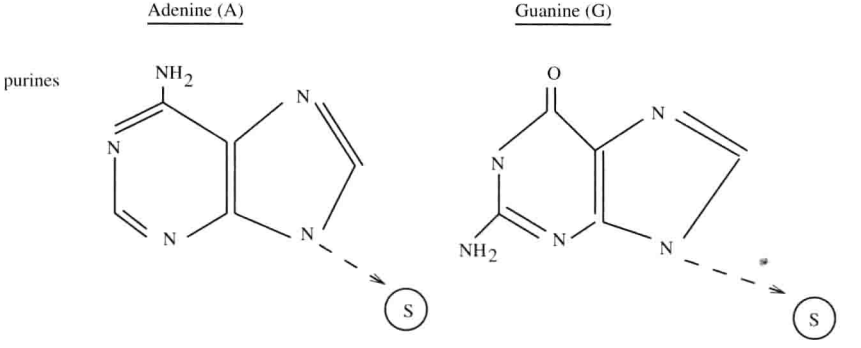


Details of DNA Backbone

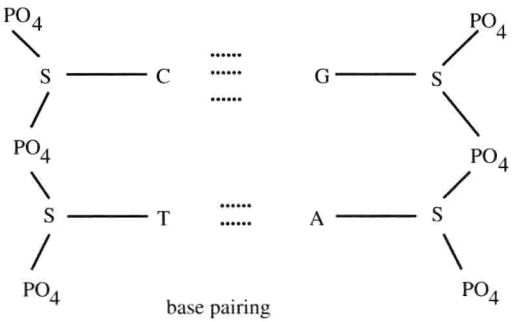
(attached to 1′ of the sugar, again with elimination of water) of four types. Two of these are *pyrimidines*, built on a six-member ring of four carbons and two nitrogens (single and double bonds are indicated, carbons are implicit at line junctions). Note: Pyrimidine, cytosine, and thymine all have the letter *y*.



Two are the more bulky *purines*, built on joined five- and six-member rings (adenine, with empirical formula $H_5C_5N_5$, used to have the threatening name pentahydrogen cyanide, of possible evolutionary significance).



DNA chains are normally present as pairs, in the famous Watson–Crick *double-helix* conformation, enhancing their mechanical integrity. The two strands are bound through pairs of bases, pyrimidines to purines, by means of *hydrogen bonds* (. . . . . .), and chemical fitting requires that A must pair with T, G with C; thus each chain uniquely determines its partner. The DNA "alphabet" consists of only the four letters A, T, G, and C, but the full text is very long indeed, some 3×10^9 base pairs in the human. Roughly 3% of *our* DNA four-letter information is allocated to genes, "words" that translate into the proteins that, among other activities, create the enzymatic machinery that drives biochemistry, as well as instructional elements, the rest having unknown – perhaps mechanical – function.

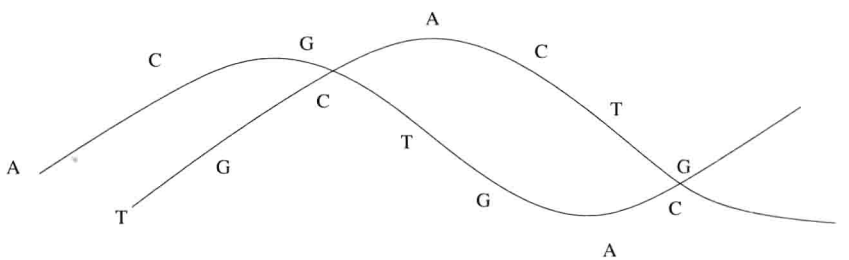


base pairing

Double-chain DNA is typically represented in linear fashion, e.g.,

$$\begin{array}{cccccccccccc} 5' & - & A & - & C & - & G & - & T & - & G & -A & - & C & - & 3' \\ & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & \vdots & & \vdots & & \\ 3' & - & T & - & G & - & C & - & A & - & C & -T & - & G & - & 5' \end{array}$$

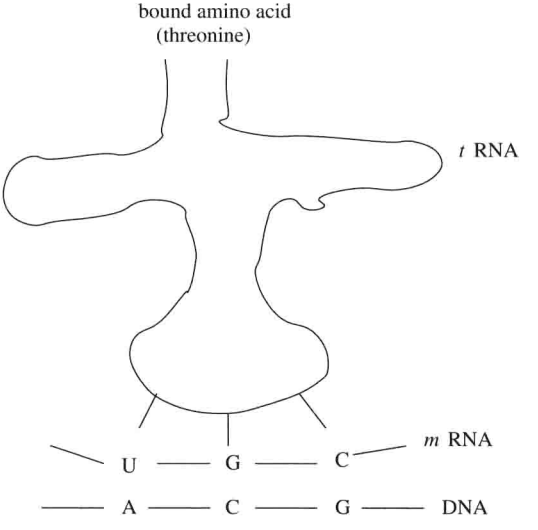
(although the unique base pairing means that say the single 5′–3′ chain suffices), but because of the offset between 3′ and 5′ positions, the spatial structure is that of a spiral ribbon.



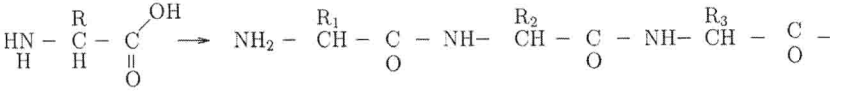
paired strands

Even the small portions of DNA – the genes – that code for proteins are not present in compact regions but, especially in the compact-nucleus eukaryotes, are interrupted by noncoding (and often highly repetitious) *introns*. The coding fragments – or *exons* – are also flanked by instructional subsequences, so that a small gene might look like: (5′) upstream enhancer, promoter, start site, exon, intron, exon, poly-A site, stop site, downstream enhancer (3′). However, the vast remaining "junk DNA" – also riddled by fairly complex repeats (ALU, 300 base pairs; L1, very long; microsatellites, very short) – aside from its obvious mechanical properties, leading, e.g., to a supercoiled structure grafted onto the double helix, is of unknown function, and may be only an evolutionary relic.

The major steps in the DNA → protein sequence are well studied. Separation of the chains allows the exon–intron gene region of one of the chains to be read or *transcribed* to a pre-RNA chain of nucleotides (similar to the duplication of DNA needed in cell division) that differs from DNA by the substitution of *U* (uracil) for the *T* of DNA and by ribose (with a 2′-OH) for deoxyribose. The introns are then spliced out (by a signal still incompletely understood) to create messenger RNA, or *m*-RNA, which almost always (RNA can also be an end product) is itself read by transfer RNA, or *t*-RNA, which *translates*



by setting up a specific amino acid for each base triplet of the *m*-RNA, or *codon* of the DNA, the amino acids then joining to form protein. The triplets code for 20 amino acids (as well as the start codon AUG at its first occurrence and stop codons UAA, UAG, UGA) when present in exons, and they come in four main varieties: nonpolar (hydrophobic), polar uncharged, + charged



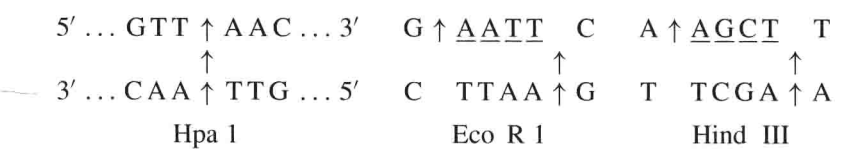
(basic), and – charged (acidic). Of course, there are always exceptions, and stop codons seem to be responsible as well for incorporation of crucial trace metals (selenium, zinc, etc.) into protein. Because there are 64 possible codons, there is a good deal of ambiguity, and the third member of the triplet

is irrelevant in most cases. As we go along a DNA double strand (5×10^6 base pairs in *E. coli*, 3×10^9 – in 46 chromosomes – for us) there are six possible "reading frames" for triplets (3 times $5' \rightarrow 3'$ for either strand), and the correct one is selected by a start signal. The three-dimensional spatial or folding structure is important for the DNA and crucial for the resulting protein, but this is determined (precisely how is only partially clear – chaperonins, large protein templates, certainly help) by the one-dimensional sequence or primary structure, which is what we focus on.

The initial information that we seek is then the identity of the sequence of $\approx 3 \times 10^9$ "letters" that, e.g., mark us as human beings, and some of whose deviations mark us as biochemically imperfect human beings. Many techniques have been suggested, and more are being suggested all the time, but almost all rely on the availability of exquisitely selective enzymes.

1.2. Restriction Fragments

Although our DNA is parceled among 46 chromosomes, (22 pairs plus 2 sex chromosomes) each is much too large to permit direct analysis. There are many ways, mechanical, enzymatic, or other, to decompose the DNA into more malleable fragments. In particular, there are (type II) *restriction enzymes* available that cut specific subsequences (usually four, six, or eight letters long) in a specific fashion (Nathans and Smith, 1975). These enzymes are used by bacteria to inactivate viral DNA, while their own are protected by methylation. They are almost all *reverse palindromes* (one, read 5′–3′, is the same as the other strand, read 3′–5′), for reasons not agreed on. In this way, we create much shorter two-strand fragments, 25–500 Kb (kilobase pairs) depending, to analyze (the loose ends can also bind other loose ends created by the same enzyme to form recombinant DNA). In practice, many copies of the DNA are made, and only a portion of the possible cuts is performed, so that a highly diverse set of overlapping fragments is produced (see Section 1.3).



The fragments, which can be replicated or cloned in various ways, can then serve as a low-resolution signature of the DNA chain, or a large segment thereof, provided that they are characterized in some fashion. Of several in current use, the oldest characterization is the restriction-enzyme *fingerprint*: the set of lengths of subfragments formed, e.g., by further enzymatic

digestion. These are standardly found, with some error, by migration in gel electrophoresis. Typically (Schaffer, 1983) we use the empirical relation $(m - m_0)(l - l_0) = c$, where m is migration distance and l is the fragment length, with m_0, l_0, and c obtained by least-squares fitting with a set of accompanying standard fragments (l_i, m_i): Define $c(m, l) = (m - m_0)(l - l_0)$ and minimize $Q = \sum_i [c(m_i, l_i) - c_{\text{av}}]^2$ to get m_0, l_0, and c estimates, and then compute by $l = l_0 + c_{\text{av}}/(m - m_0)$. What size fragments do we expect so that we can design suitable experiments? This is not as trivial as it sounds and will give us some idea of the thought processes we may be called on to supply (Waterman, 1983). A heuristic approach (Lander, 1989) will suffice for now.

It is sufficient to concentrate on one strand, as the other supplies no further information. Suppose the one-enzyme cut signal is a six-letter "word," (5′) $b_1\, b_2\, b_3\, b_4\, b_5\, b_6$ (3′), and, as a zeroth-order approximation to the statistics of DNA, imagine that the letters occur independently and with equal probability, $p(A) = p(C) = p(T) = p(G) = 1/4$, at each site. Then, for each site, the probability of starting and completing the word to the right is simply $\frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4}$,

$$p(b_1 b_2 b_3 b_4 b_5 b_6) = 1/4^6.$$

Suppose we have found one word and continue down the strand looking for the next occurrence. Assuming that $b_1 b_2 b_3 b_4 b_5 b_6$ cannot initiate a displaced version of itself, e.g., $b_5 b_6 \neq b_1 b_2$, we start after the word ends. Then the probability of not seeing a new word start for $l - 1$ moves but seeing one at the lth move is clearly the *geometric* distribution

$$p(l) = (1 - 1/4^6)^{l-1}\, 1/4^6$$

{or, because $1/4^6$ is very small, $p(l) \sim [(1/4^6)e^{-l/4^6}]$, the continuous *exponential* distribution}. The mean distance to the next word is then the mathematical expectation

$$\mu = E(l) = \sum_{l=0}^{\infty} \frac{1}{4^6} \left(1 - \frac{1}{4^6}\right)^{l-1} l.$$

On evaluation, $[\sum_{l=0}^{\infty} \alpha l (1 - \alpha)^{l-1} = -\alpha \frac{\partial}{\partial \alpha} \sum_{l=0}^{\infty} (1 - \alpha)^l = -\alpha \frac{\partial}{\partial \alpha} \frac{1}{\alpha} = \frac{1}{\alpha}]$, we have

$$\mu(b_1 b_2 b_3 b_4 b_5 b_6) = 4^6 = 4096.$$

The preceding argument will not hold for self-overlapping words, as the absence of a word starting at a given site slightly biases the possibilities for

words starting at the next six sites, but because p is so small, this correlation effect is very small. We also have to distinguish between allowing two occurrences to overlap and not allowing it. In fact, a careful mathematical analysis (Guibas and Odlyzko, 1980) shows that the relation

$$\mu = 1/P$$

holds exactly for a long *renewal process*, one in which all the letters of a word are removed before we start counting again; here μ is the mean repeat distance from the beginning of the pattern and P is the probability that a renewal starts at a given site. Interestingly, this is precisely the situation that is said to exist with restriction enzymes – for a recognition site such as TAG CTA with self-overlap after moving four bases, a subsequence TAGCTAGCTA would be cut only once, whatever the direction of travel of the enzyme – there would not be enough left to cut a second time (the main reason seems to be that an enzyme needs something to hold onto and cannot work directly on a cut end). If this is the case, the mean repeat distance will change. In this example, we still have the basic $p(\text{TAGCTA}) = 1/4^6$, but the unrestricted p at site n is composed of either a repeat, say at site n, or a repeat at site $n-4$, followed by the occurrence of GCTA to complete the TA pair: $p = P + 4^{-4}P$. Hence $\mu = 1/P = (1 + 4^{-4})/p = 4^6 + 4^2 = 4112$. More generally, we find

$$\mu = 4^6(1 + e_1/4 + \cdots + e_5/4^5),$$

where $e_i = 1$ for an overlap at a shift by i sites, otherwise $e_i = 0$.

The relevance of the above discussion in practice is certainly marginal, as the significance of such deviations is restricted to very short fragments, which are generally not detected anyway. However, the assumption of independent equal probabilities of bases is another story. To start with, these probabilities depend on the organism and the part of the genome in question, so that we should really write instead

$$p(b_1 \cdots b_6) = p(b_1) \cdots p(b_6),$$

and this can make a considerable difference, which is observed. To continue, we need not have the independence $p(bb') = p(b)\,p(b')$; rather,

$$g(bb') = p(b\,b') \,/\, p(b)\,p(b')$$

measures the correlation of successive bases – it is as low as $g(CG) \sim 0.4$. If this successive pair correlation or Markov chain effect is the only correlation present, we would then have

$$p(b_1 \cdots b_6) = p(b_1) \cdots p(b_6)\; g(b_1 b_2)\, g(b_2 b_3)\, g(b_3\, b_4)\, g(b_4\, b_5)\, g(b_5\, b_6),$$