BIOINFORMATICS
Electronic edition http://www.bioinformatics.oupjournals.org

VOLUME 18
NUMBER 8
AUGUST 2002
PAGES 1021–1033

Why are complementary DNA strands symmetric?

Pierre-François Baisnée[1], *Steve Hampson*[1] *and Pierre Baldi*[1,2,*]

[1]*Department of Information and Computer Science, Institute for Genomics and Bioinformatics*

[2]*Department of Biological Chemistry, College of Medicine, University of California, Irvine, CA 92697-3425, USA*

Received on November 24, 2001; revised on March 7, 2002; accepted on March 11, 2002.

* To whom correspondence should be addressed.

GO BACK
CLOSE FILE

Abstract

Motivation: *Over sufficiently long windows, complementary strands of DNA tend to have the same base composition. A few reports have indicated that this first-order parity rule extends at higher orders to oligonucleotide composition, at least in some organisms or taxa. However, the scientific literature falls short of providing a comprehensive study of reverse-complement symmetry at multiple orders and across the kingdom of life. It also lacks a characterization of this symmetry and a convincing explanation or clarification of its origin.*

Results: *We develop methods to measure and characterize symmetry at multiple orders, and analyze a wide set of genomes, encompassing single- and double-stranded RNA and DNA viruses, bacteria, archae, mitochondria, and eukaryota. We quantify symmetry at orders 1 to 9 for contiguous sequences and pools of coding and non-coding upstream regions, compare the observed symmetry levels to those predicted by simple statistical models, and factor out the effect of lower-order distributions. We establish the universality and variability range of first-order strand symmetry, as well as of its higher-order extensions, and demonstrate the existence of genuine high-order symmetric constraints. We show that ubiquitous reverse-complement symmetry does not result from a single cause, such as point mutation or recombination, but rather emerges from the combined effects of a wide spectrum of mechanisms operating at multiple orders and length scales.*

Contact: *baisnee@ics.uci.edu; hampson@ics.uci.edu; pfbaldi@ics.uci.edu*

Data: *http://promoter.ics.uci.edu/RevCompSym/*

Introduction

Chargaff's famous first parity rule (Chargaff, 1951) states that, in any piece of (double-stranded) DNA, the number of As exactly equals the number of Ts, and the number of Cs exactly equals the number of Gs. The Watson and Crick base-pairing model fully explains this property of double-stranded DNA molecules. It is less widely known that the first parity rule approximately holds true within *single* DNA strands, over windows of sufficient size, often in the order of 1000 bp (Bell and Forsdyke, 1999a; Fickett *et al.*, 1992; Forsdyke, 1995a). This intra-strand parity rule can equivalently be stated from a double-strand perspective: complementary DNA strands tend to have the same base composition and are in this respect symmetric. Previous reports have indicated that the rule extends from the first order (base composition) to higher orders (oligonucleotide composition), at least in some organisms or taxa (Forsdyke, 1995a; Hampson *et al.*, 2000; Prabhu, 1993). At the second order, for instance, the dinucleotide CT would accordingly tend to be equi-frequent in reverse-complementary strands, or as frequent as its reverse-complement AG within a strand.

The intra-strand parity of complementary bases is sometimes called 'Chargaff's second parity rule' (Bell and Forsdyke, 1999a,b; Forsdyke and Mortimer, 2000). However, a careful reading of Chargaff's papers reveals that the only intra-strand parity he reports is that of 6-amino (A+C) and 6-oxo (G+T) compounds (Chargaff, 1951, 1979; Karkas *et al.*, 1968, 1970; Lin and Chargaff, 1967; Magasanik and Chargaff, 1989; Rudner *et al.*, 1968a,b,

Abstract
Introduction
Materials and methods
Results
Discussion
Acknowledgements
References
Appendix
GO BACK
CLOSE FILE

1969). Furthermore, a review of the scientific literature shows that DNA strand symmetry is not the well established and explained fact that such a prestigious but apparently erroneous paternity may suggest. The literature specifically addressing reverse-complement symmetry indeed consists of a few isolated reports, each suffering from at least one, if not all, of the following shortcomings: (a) limited set of sequences; (b) analysis at low orders only; (c) purely qualitative results, flawed quantitative analysis, or lack of characterization; (d) absence of explanation or unconvincing explanation.

In contrast, there is an abundant literature on first-order asymmetries, known as ‘skews’. Typically amounting to 4% (see **Methods** for an explanation of symmetry and asymmetry measures), skews develop locally in many prokaryotes, viruses and mitochondria, sometimes extending over very long stretches. They are independent of AT content, but correlate with the direction of replication and/or gene orientation, and are likely to result from mutation biases related to the functions born by each strand—leading or lagging with respect to replication, template or synonymous with respect to transcription. G is generally found in excess on synonymous strands and on leading strands; the same is often true of T, although to a lesser extent (Blattner *et al.*, 1997; Burland *et al.*, 1993; Frank and Lobry, 1999; Grigoriev, 1999; Kano-Sueoka *et al.*, 1999; Lobry, 1996a,b; Perna and Kocher, 1995; Sueoka, 1995; Tillier and Collins, 2000; Wu and Maeda, 1987). Skews are consequently routinely used for the prediction of replication origins, and skew plots now figure in most genome analyses. However, while it provides convincing explanations for local asymmetry, the

GO BACK
CLOSE FILE

literature apparently takes first-order strand symmetry for granted in the first place. The underlying assumption is that base-composition symmetry results from single-point mutations that equally affect complementary strands, as demonstrated in the case of simple models of DNA evolution (Lobry, 1995; Lobry and Lobry, 1999).

In addition, high-order symmetry is widely considered, implicitly or explicitly, as the consequence of first-order symmetry. In principle, the latter might indeed induce the former, given that combinations of nucleotides randomly drawn from a symmetric pool are likely to result in symmetric oligonucleotide distributions (e.g. if $P(XY) = P(X)P(Y)$, and $P(A) = P(T)$, then $P(AA) = P(A)^2 = P(T)^2 = P(TT)$). Two high-order symmetry mechanisms, however, have been suggested. Fickett *et al.* (1992) noted that strand inversion—resulting from recombination events in which fragments of complementary strands are swapped—could be an explanation. From a more speculative perspective, Forsdyke (1995a,b) suggested that the selection of stem-loop structures might be a primary source of symmetry. Since stem-loop formation relies on base pairing of nucleotides on the same strand in the stem region, their selection would induce Nmer reverse-complement symmetry up to the length of the stems. To the best of our knowledge, however, no attempt has been made to determine whether high-order symmetry could result entirely from first-order parity, to investigate whether explanations relying on a single factor are adequate, or to characterize the nature or clarify the origin of strand symmetry in any other way.

Here, we develop methods to quantify symmetry at all orders and to assess whether high-order parities result from those existing at lower orders. We first establish the universality and variability of symmetry at orders 1 to 9 across a large set of genomes ranging from viruses to organelles, to higher eukaryotes. We then demonstrate the existence of genuine high-order symmetries that do not entirely result from lower-order ones, and invalidate explanations relying on a single mechanism—be it single-point mutation at the first order, or recombination events resulting in strand inversion. We show that symmetry instead results from an array of mechanisms operating at multiple orders and imprinting DNA sequences at different length-scales.

Materials and methods

Symmetry: distributions and plots

To study reverse-complement symmetry at a given order N, we count all overlapping occurrences of each of the 4^N possible oligonucleotides of length N, along a given DNA strand and over a given length, in $5'$ to $3'$ orientation. We thus obtain a distribution of the form $\pi(X_1 \ldots X_N)$, where $X_1 \ldots X_N$ represents Nmers. Similar results are obtained with non-overlapping counts. Perfect strand symmetry of order N occurs when every Nmer is as frequent as its reverse complement, i.e. when $f(X_1 \ldots X_N) = f(\bar{X}_N \ldots \bar{X}_1)$. Since the size of the distribution grows exponentially with N, large data sets are needed at high orders to ensure that most Nmers are represented and that their relative abundance is accurately estimated.

To visually assess the reverse-complement symmetry of a sequence, we plot the frequencies (or counts) observed on one strand against those observed on the complementary strand—which is obviously virtual in the case of single-stranded genomes. Such plots are necessarily symmetric with respect to the diagonal line, reflecting strand reverse-complementarity. Points aligned on the diagonal, however, reveal perfect strand symmetry, while points that are distant from the diagonal reveal strand asymmetry.

Abstract
Introduction
Materials and methods
Results
Discussion
Acknowledgements
References
Appendix
GO BACK
CLOSE FILE

Symmetry and similarity measures

We measure the Nth-order strand symmetry of any given sequence as the similarity between its Nmer distribution f and the Nmer distribution f' of its actual or virtual reverse-complement. We thus derived indices from standard distance or divergence measures such as the L^p distances or the Kullback–Liebler relative entropy or divergence (Baldi and Brunak, 2001). In practice, we use: (a) an index based on the L^1 distance, i.e. the sum of the absolute values of the differences between oligonucleotide frequencies:

$$S^1 = 1 - \frac{\sum_i |f_i - f_i'|}{\sum_i |f_i| + |f_i'|}; \tag{1}$$

or (b) Pearson's linear correlation coefficient:

$$S^C = C(f, f'). \tag{2}$$

Both indices can be computed over complete sets of Nmers or on particular subsets. When N is even, for instance, there are $4^{N/2}$ Nmers that are identical to their reverse-complement, which can significantly increase symmetry measures at low orders N. Such reverse-complement invariant Nmers can either be taken into account to compute the overall symmetry level of the full distribution, or discarded in order to capture the specific symmetry level of its non-palindromic subset. Except when otherwise specified, we here measure symmetry on full distributions, in which case the denominator of (Equation 1) is equal to 2.

Note that one can also use S^1 and S^C to measure the strand symmetry of the discrepancies $f = o - r$ or $f = o/r$ between an observed Nmer distribution o and any reference Nth-order distribution r.

S^1 ranges from 0 (asymmetry/dissimilarity) to 1 (perfect symmetry/similarity). When computed on distributions, it represents the percentage of Nmer occurrences that are symmetrically distributed among complementary strands. It generalizes classical measures of first-order asymmetries—AT and GC skews—which are expressed as (A-T)/(A+T) and (G-C)/(G+C). Its complement to 1 (an asymmetry index) indeed corresponds to the weighted average of the absolute values of the skews of reverse-complementary bases or Nmers. This is easily seen by comparing (Equation 1) to the weighted sum of the absolute values of the skews $(f_i - f_i')/(f_i + f_i')$, with weights $(f_i + f_i')/\sum_i(f_i + f_i')$. In the **Appendix**, we show that, for any given sequence, S^1 monotonically decreases as N increases.

S^C ranges from -1 to 1 and generally yields results that are qualitatively similar to those obtained with S^1. However, there are a number of differences, and some precautions need to be taken when using S^C. First, it is well known that correlation is sensitive to outliers. In genomic sequences, over-represented Nmers, such as runs of As or Ts (poly(A) and poly(T) tracts), can in particular bias the S^C symmetry index. More generally, for a constant S^1 level, sequences displaying a more widespread Nmer distribution tend to result in higher S^C values. Second, a perfectly uniform base or oligonucleotide distribution of order N—for which $f_i = 1/4^N$—satisfies $S^1 = 1$ and therefore is perfectly symmetric according to this index. However, S^C is not defined in this case, and small random perturbations around uniformity can cause low-order S^C values to fluctuate widely. Third, for sequences that are too short to measure Nth-order statistics, Nmers that are not represented can increase S^C when taken into account, especially at low orders. Lastly, unlike S^1, S^C does not necessarily monotonically decrease as N increases.

We also use S^1 and S^C in sliding windows of varying sizes to measure the evenness of the distribution of genes or other features among complementary strands, both in terms of their number of occurrences and in terms of their base-pair coverage.

Statistical models of symmetry

To further assess strand symmetry and gain insight on its origin, it is useful to build strand-symmetric statistical models of DNA sequences, and compare the symmetry level they predict to the levels observed in actual DNA sequences. For this purpose, we model biological sequences using Markov models. A DNA or RNA Markov model of order N has 4^N parameters associated with the transition probabilities $P(X_N|X_1 \dots X_{N-1}) = P(X_1 \dots X_N)/P(X_1 \dots X_{N-1})$, for all possible $X_1 \dots X_N$ in the alphabet, together with a starting distribution of the form $\pi(X_1 \dots X_{N-1})$. Since the number of parameters grows exponentially, only models up to a certain order can be determined from a finite data set.

We built models of order 1 to 9 that matched a wide spectrum of highly symmetric biological distributions, and then slightly modified the

Abstract
Introduction
Materials and methods
Results
Discussion
Acknowledgements
References
Appendix
GO BACK
CLOSE FILE

model parameters to force perfect reverse-complement symmetry. Through computer simulations, we then generated DNA sequences of various lengths and estimated the expectation (μ) and variance (σ^2) of S^1 at orders 1 to 9. For uniform distributions, we show that $\mu \approx 1 - \sqrt{4^N - 1}/\sqrt{\pi L}$ and that $\sigma^2 \approx (\pi - 2)(4^N - 1)/(2\pi 4^N L)$, where L and N respectively represent the sequence length and the order at which symmetry is measured (see **Appendix**). These estimations are in good agreement with computer simulations.

Qi and Cuticchia (2001) have tested the significance of symmetry using a paired Student's t-test. However, the corresponding null hypothesis, i.e. the fact that the mean difference of counts of Nmers and their reverse complement is zero, is likely to be also verified for asymmetric distributions and this test is irrelevant. A χ^2 test on the distributions of complementary strands could be used instead. However, biological symmetry is not perfect, and even for the most symmetric biological sequences, complementary strand would have significantly different distributions according to such a test. The analytical approximations and the simulations we use to estimate the distribution of symmetry measures are consequently more appropriate approaches.

Restrictions and extensions

A distribution of order N imposes constraints over lower-order Mmers ($M < N$) and thus induces a *unique* distribution of order M called the restriction or projection of the original distribution. This projection can be calculated using

$P(X_1 \dots X_M) = \sum_{Y_{M+1} \dots Y_N} P(X_1 \dots X_M Y_{M+1} \dots Y_N)$. On the other hand, a distribution of order M can have multiple extensions to a distribution of order N, $N > M$. A given distribution of order M, however, yields a unique *factorial extension* or predicted distribution at any order $N > M$. For instance, a first-order distribution defined by the parameters p_X (p_A, p_C, p_G, p_T) has a second-order factorial extension with parameters $p_{XY} = p_X p_Y$.

Alternatively, one can estimate the factorial extension or the restriction of a distribution of order N at any order O by generating a sufficiently long random string with the Markov model of order N, and by computing its statistics of order O.

When a distribution of order O is symmetric ($S^1_O = 1$), then: (a) its unique restriction to any lower order M is also symmetric; (b) its multiple extensions to any higher order N need not be symmetric; (c) its unique factorial extension to any order N, however, is also symmetric.

Evidencing high-order constraints

To put in evidence high-order constraints, we measure the discrepancies that arise at any order N between the factorial extensions of lower-order distributions of order M and the actual Nth-order distribution.

For this purpose, we use the distance $1 - S^1$. Alternatively, we use S^C as a measure of the fit between predicted and observed distributions. Discrepancies necessarily result from biological mechanisms that operate above the Mth order

Abstract
Introduction
Materials and methods
Results
Discussion
Acknowledgements
References
Appendix
GO BACK
CLOSE FILE

and tend to select or exclude Kmers with $K > M$. Such distance or fit measures, however, do not reflect the relative importance of the mechanisms that shape DNA sequences at various orders.

Residual symmetry

To assess the symmetry of high-order constraints, we factor out at any order N the effect of any lower-order distribution of order M, for each oligonucleotide and on each strand. This is achieved by quantifying the discrepancies between observed Nmer frequencies (o_i) and the corresponding expected frequencies (e_i) according to the factorial extension of the distribution of order M. We thus form ratios ($f_i = o_i/e_i$) or differences ($f_i = o_i - e_i$), which both yield qualitatively similar results. If one of the denominators happens to be zero, small pseudo-counts equivalent to Dirichlet priors (Baldi and Brunak, 2001) can be used to avoid infinite ratios. We then compute S^1 and S^C on these ratios or differences, thus measuring *residual symmetry* that results from genuine high-order ($> M$) constraints. It can be shown that S^1 residual symmetry drops to a value close to 0.3 (instead of 0 for S^C) after removal of all symmetry (see **Appendix**).

Data

We analyzed 396 sequences representing full or partial chromosomes and genomes of 6 eukaryota, 11 bacteria, 10 archaea, 100 single- or double-stranded

RNA or DNA viruses and 192 mitochondria. Most sequences were downloaded from GenBank (Benson *et al.*, 2000) or Entrez (Schuler *et al.*, 1996). Our sample of mitochondria includes every complete genome that was available as of May 26, 2001.

Strand symmetry is a property of individual DNA molecules and ought to be measured in contiguous sequences corresponding to a specific strand. Pooling sequences that belong to complementary strands or to different chromosomes can artificially increase any degree of symmetry. In this respect, it is worth noting that some published eukaryotic chromosome sequences are not yet fully oriented: some of their sub-sequences, surrounded by gaps, may belong to one strand or the other. We analyzed such sequences, as well as their largest fully oriented sub-sequences. We also compared fully oriented releases with earlier, partially oriented ones. Although we concluded that mis-oriented sub-sequences only moderately affect the overall symmetry level, we discarded incompletely oriented published chromosome sequences from our data set. Instead, we used large, fully-oriented sub-sequences, or the yet-unpublished most recent oriented sequences available from the sequencing and assembly centers. See http://promoter.ics.uci.edu/RevCompSym/ for details and a complete list of sequences.

While strand symmetry must be measured over a specific strand, its origin can be investigated by pooling non-contiguous regions that share a given characteristic, such as their leading or lagging status during replication, their coding or non-coding nature, their position upstream of genes, etc. The specific symmetry levels of such pools may provide insight on the mechanisms that promote or disrupt symmetry. When two pools gathered on complementary strand have similar compositions, it is reasonable to merge them to study some aspects of symmetry with a higher statistical accuracy. The same applies to pools extracted from different chromosomes.

Here, within eukaryotic genomes, we pooled and analyzed separately all identified coding regions and their 500 bp long upstream non-coding regions, discarding non-coding regions that overlapped on opposite strands, and coding sequences that overlapped with the 500 bp long upstream region of a coding sequence on the opposite strand.

Lastly, to further investigate how sequence length affects symmetry levels, we used two complementary methods: (a) iterative halving of the data; (b) sliding windows.

Table 1. Single-stranded base composition (%) of yeast nuclear and mitochondrial chromosomes. The corresponding AT and GC skews, and the overall S^1 and S^C symmetry levels are indicated

	size	A (%)	T (%)	AT skew (%)	G (%)	C (%)	GC skew (%)	S^1 (%)	S^C
Chr. 1	230 203	30.33	30.39	−0.10	19.88	19.39	1.24	99.45	0.9979
Chr. 2	813 140	30.70	30.95	−0.41	18.99	19.36	−0.97	99.37	0.9985
Chr. 3	315 339	31.14	30.30	1.37	18.85	19.70	−2.21	98.31	0.9891
Chr. 4	1 531 929	31.12	30.97	0.24	19.02	18.89	0.35	99.72	0.9997
Chr. 5	576 870	30.60	30.89	−0.47	19.47	19.04	1.12	99.28	0.9980
Chr. 6	270 148	30.70	30.57	0.21	19.41	19.32	0.22	99.79	0.9998
Chr. 7	1 090 936	31.01	30.93	0.14	19.02	19.04	−0.08	99.89	0.9999
Chr. 8	562 638	30.93	30.58	0.57	19.10	19.39	−0.74	99.36	0.9984
Chr. 9	439 885	30.54	30.56	−0.03	19.47	19.43	0.12	99.94	1.0000
Chr. 10	745 440	31.00	30.63	0.60	19.29	19.08	0.56	99.42	0.9986
Chr. 11	666 445	30.92	31.01	−0.15	18.91	19.16	−0.67	99.65	0.9995
Chr. 12	1 078 172	30.66	30.86	−0.33	19.21	19.27	−0.17	99.73	0.9997
Chr. 13	924 430	30.97	30.83	0.23	19.09	19.12	−0.09	99.82	0.9998
Chr. 14	784 328	30.80	30.56	0.38	19.30	19.34	−0.09	99.73	0.9996
Chr. 15	1 091 283	31.10	30.74	0.58	19.01	19.15	−0.39	99.49	0.9989
Chr. 16	948 061	31.01	30.93	0.12	19.04	19.02	0.04	99.91	1.0000
Chr. mt	85 779	42.17	40.73	1.74	9.11	8.00	6.47	97.45	0.9970

Results

Example of yeast chromosomes

To illustrate DNA strand symmetry, Tables 1 and 2 show the single-stranded nucleotide and dinucleotide composition of the 16 nuclear chromosomes and the mitochondrial chromosome of *S. cerevisiae*. Remarkably, the number of As

Table 2. Single-stranded dinucleotide composition (%) and S^1 and S^C symmetry levels of yeast nuclear and mitochondrial chromosomes. Reverse-complement invariant dinucleotides AT, TA, CG and GC are excluded from the table and S^1 and S^C calculation

	AA	TT	AC	GT	AG	CT	CA	TG	CC	GG	GA	TC	S^1	S^C
Chr. 1	10.4	10.5	5.4	5.6	5.9	5.7	6.6	6.8	4.0	4.1	6.3	6.1	98.75	0.9964
Chr. 2	10.7	10.9	5.3	5.2	5.8	5.9	6.6	6.4	4.0	3.9	6.1	6.3	98.93	0.9981
Chr. 3	10.9	10.3	5.6	5.2	5.7	5.8	6.8	6.3	4.1	3.8	6.1	6.3	97.33	0.9914
Chr. 4	10.9	10.9	5.2	5.2	5.9	5.8	6.5	6.5	3.8	3.8	6.3	6.2	99.61	0.9997
Chr. 5	10.6	10.8	5.2	5.4	5.8	5.8	6.4	6.6	3.9	4.0	6.2	6.1	98.86	0.9982
Chr. 6	10.7	10.6	5.3	5.3	5.8	5.9	6.4	6.6	4.0	4.0	6.3	6.2	99.37	0.9993
Chr. 7	10.9	10.9	5.3	5.2	5.8	5.8	6.5	6.4	3.9	3.8	6.2	6.2	99.63	0.9999
Chr. 8	10.9	10.6	5.4	5.3	5.8	5.8	6.6	6.5	4.0	3.9	6.2	6.2	99.04	0.9990
Chr. 9	10.6	10.6	5.3	5.4	5.9	5.9	6.5	6.5	4.0	4.0	6.2	6.2	99.80	0.9999
Chr. 10	10.9	10.6	5.3	5.3	5.9	5.8	6.5	6.5	3.8	3.9	6.3	6.2	99.17	0.9997
Chr. 11	10.9	10.9	5.2	5.2	5.8	5.9	6.5	6.4	3.9	3.8	6.2	6.3	99.45	0.9992
Chr. 12	10.7	10.8	5.2	5.3	5.9	5.9	6.5	6.5	4.0	3.9	6.3	6.3	99.55	0.9999
Chr. 13	10.9	10.8	5.3	5.3	5.8	5.8	6.5	6.4	3.9	3.9	6.2	6.2	99.68	1.0000
Chr. 14	10.7	10.6	5.4	5.3	5.9	5.8	6.5	6.5	4.0	3.9	6.3	6.2	99.52	0.9999
Chr. 15	11.0	10.8	5.3	5.2	5.8	5.8	6.5	6.4	3.9	3.9	6.2	6.2	99.26	0.9998
Chr. 16	10.9	10.9	5.2	5.2	5.9	5.9	6.4	6.4	3.9	3.8	6.3	6.2	99.79	1.0000
Chr. mt	16.0	14.7	2.1	2.6	3.0	2.5	2.1	2.3	2.3	2.6	3.0	2.6	94.28	0.9975

and Ts, or the number of Cs and Gs, is approximately the same when counted along any single strand. This translates in relatively low AT and GC skews and high overall S^1 and S^C symmetry levels. In addition, this reverse-complement parity holds also for dinucleotides (e.g. $f(AC) \approx f(GT)$). While nuclear chromosomes are homogeneous, the mitochondrial chromosome displays a different composition and a lower symmetry. The slight discrepancies in

Abstract
Introduction
Materials and methods
Results
Discussion
Acknowledgements
References
Appendix
GO BACK
CLOSE FILE

chromosome ranking according to S^1 and S^C illustrate the sensitivity of S^C to the spread of Nmer distributions (**Methods**). In particular, the high S^C values observed for the mitochondrial chromosome result from a highly symmetric outlier (poly-A and poly-T tracts).

Base composition symmetry

Our analysis of genomic sequences reveals that first-order symmetry steadily increases with DNA length, both across organisms (Fig. 1) and within genomes (not shown). The increase is linear in a (log(length), log(1-S^1)) space. Beyond 10^5 bp, S^1 symmetry generally exceeds 99%. This is the case for every eukaryotic, archaeal, and bacterial complete chromosome we examined, with the exception of *M. pneumoniae*, yeast chromosome 3, and two chromosome fragments of *D. melanogaster* (respectively 1.16%, 1.69%, 1.21% and 1.50% asymmetry levels, corresponding to the lowest bacteria and eukaryota points in Fig. 1). Some double-stranded DNA viruses also display high symmetry levels ($>$ 99%). Large mitochondrial chromosomes can be highly symmetric, whereas smaller ones range from moderate to very high asymmetry levels (1% to 48%). Smaller genomes of single-stranded DNA viruses, RNA viruses, retroids and extra-chromosomal elements all display moderate to high levels of asymmetry (1% to 26%).

For comparison purposes, we also plot in Fig. 1 the average base-composition symmetry level that would be expected when generating DNA sequences by randomly drawing nucleotides from a uniformly distributed (A,C,G,T) pool, along with the level found 3 standard deviations below the expectation (upper solid and dotted lines in Fig. 1). This uniform model yields approximately the same results as any other symmetric first-order one, and thus provides us with an estimation of the maximum average symmetry and the minimum variability that can be expected for biological sequences. While at all lengths the most symmetric sequences approach or exceed such maxima, most points are widely spread along the (logarithmic) y-axis and are found more than 3 standard deviations below. The average level that a perfectly symmetric first-order model yields at any given length L is reached by biological sequences one to two orders of magnitude longer.

Higher-order symmetric Markov models (not shown for the sake of readability) yield lower expectations and higher standard deviations, resulting in lines that are parallel to those plotted for a first-order model. The corresponding symmetry levels are still higher and less variable than those observed in biological sequences.

However, symmetry values are confined to a diagonal band, and exceed, for instance, the average symmetry level that a perfectly symmetric first-order model would yield for heptanucleotides (lower line in Fig. 1). Furthermore, a visual inspection reveals that, in logarithmic coordinates, the distance and spread of biological levels with respect to our reference first-order model are roughly similar at all sequence lengths and across all taxa, with the notable exception of mitochondria. In Table 3, categories and taxa are ranked from most to least symmetric, according to their average normalized distance from the first-order model, i.e. after factoring out sequence length effects. While a detailed statistical analysis would require a separate study, t-tests comparing such distances among groups bring a crude confirmation to the visual observation. Mitochondria are significantly less symmetric than any of the other categories. In addition, single-stranded RNA viruses and retroids, while undistinguishable among themselves, appear to be significantly more symmetric than mitochondria, and less symmetric than the most symmetric groups.

The fact that symmetry levels can be bounded using symmetric models, that symmetry increases in a consistent manner with sequence length both across and within genomes, and that symmetry levels are roughly similarly distributed at all lengths, shows that strand symmetry can be considered as a general emerging property of large poly-nucleotide molecules submitted to evolutionary pressures. The fact that symmetry levels are lower and more variable than predicted by simple models confirms that more or less pronounced asymmetries can develop locally, and shows that: (a) first-order mechanisms, if they contribute to symmetry, do so in a variable and relatively weak manner; (b) symmetry might at least partly result from high-order mechanisms.

High-order symmetry

Remarkably, strand symmetry extends to high orders. As illustrated in Fig. 2, counts of oligonucleotides of length 2 to 9 are very similar in complementary strands and yield high symmetry values for long sequences.

Table 3. Ranking of groups or taxa after factoring out sequence length effects. Categories are reported by decreasing average symmetry order, according to the average of their normalized distances from a perfectly symmetric first-order model. Z-scores were computed on the logarithm of the asymmetry levels; their average, minimum, maximum and standard deviation are reported in successive columns for each category. Categories are labeled as in Fig. 1

Category	Size	Mean	Min	Max	Standard deviation
dD	10	2.7	−2.5	6.1	2.3
E	50	3.1	−1.9	6.3	1.7
X	27	3.4	−2.0	7.2	2.2
sD	30	3.4	0.3	6.1	1.5
B	12	3.6	1.1	5.9	1.4
dR	3	3.9	2.3	5.6	1.6
A	10	4.2	1.3	5.9	1.3
sR+	33	4.3	0.8	7.2	1.7
sR−	14	4.7	1.1	6.6	1.4
R	15	4.9	2.4	7.6	1.6
M	192	6.7	0.4	9.5	1.6

Beyond the fact that S^1 monotonically decreases with the order at which symmetry is measured (**Appendix**), long sequences are necessary to accurately measure high-order statistics and the corresponding symmetry. S^1 and S^C levels can thus drop sharply at high orders in short sequences. Two complementary empirical methods can however be used to assess whether low measures reflect inaccurate statistics or an actual asymmetry of the strand under study: (a) if symmetry levels are stable when halving the sample size or (b) if symmetry levels at neighboring orders are similar, then they are close to their asymptotic level and are accurate.

While Fig. 2 is derived from the longest sequence in our set, and one of the most symmetric, an extensive analysis reveals that, statistical accuracy problems put apart, high-order results strictly parallel first-order ones, and translate into plots similar to Fig. 1. At all orders 2 to 9, there is a general trend for symmetry to increase with sequence size. Perfect symmetry is closely approached at orders 2 to 5 in the case of the largest DNA sequences, such as human chromosome 22 ($S^1 \geq 99.6\%$).

Evidence for high-order mechanisms

Two opposite simple explanations are possible in the face of these results, both stemming from the view (Forsdyke, 1995a) that high-order symmetries are the sole consequence of first-order symmetry, or vice-versa. At one extreme, first-order mechanisms are considered as the sole cause for symmetry. To test this explanation, we systematically quantify the discrepancies that arise in complementary strands between the observed Nmer distributions and the factorial Nth-order distributions that Mmer frequencies ($M < N$) would yield in the absence of higher-order constraints (**Methods**). As exemplified in Fig. 3, discrepancies are significant and *strand-symmetric*. The distance between observed and predicted distributions increases as the orders N and M fall farther apart (Fig. 4a). The residual symmetry levels that are measured after factoring out the effect of lower-order distributions are high, even when N and M are close and the discrepancy between predicted and observed distributions are therefore low. At any order N, the residual symmetries are generally lower than the Nth-order symmetry itself. However, they consistently increase with sequence size, and again approach perfect symmetry up to the fifth order for the largest DNA sequences (Fig. 4b). Constraints operating above the predicting order M therefore tend to equally affect complementary strands. High-order phenomena, up to at least order 9, thus unambiguously contribute to strand symmetry.

Evidence for low-order mechanisms

At the opposite extreme, symmetries of all orders can be viewed as predominantly resulting from an even repartition of large-scale homogeneous features, such as coding and non-coding regions, or leading and lagging replication strands. Even if they were highly asymmetric, such features would promote symmetry at all orders when evenly distributed among complementary strands. While we cannot factor out high-order constraints to directly observe the imprint of low-order ones, we can, nonetheless, rule out this hypothesis based on a number of observations.

First, large-scale phenomena only induce symmetry over large sequences, and cannot account for the substantial symmetry that fragments a few hundred bp long often display Fickett *et al.* (1992); Forsdyke (1995a); Bell and Forsdyke (1999a)

Second, single-point mutations directly promote symmetry at the first order when they are blind with respect to strandedness (Lobry, 1995; Lobry and Lobry, 1999), or an often-moderate asymmetry when they are not (Blattner *et al.*, 1997; Burland *et al.*, 1993; Frank and Lobry, 1999; Grigoriev, 1999; Kano-Sueoka *et al.*, 1999; Lobry, 1996a,b; Perna and Kocher, 1995; Sueoka, 1995; Tillier and Collins, 2000; Wu and Maeda, 1987).

Beyond the first order, strand-symmetric selection and exclusion of oligonucleotides are plausible partial explanations for residual symmetries. Species-specific patterns of dinucleotide and trinucleotide over- and under-representation, which are consistently imprinted in sequences on the order of 50 kbp and thus constitute ‘genome signatures’, have been partly attributed to pressures directly exerted on the oligonucleotides themselves (Burge *et al.*, 1992; Campbell *et al.*, 1999; Karlin and Burge, 1995; Karlin and Mrazek, 1997; Mrazek and Karlin, 1998). It has furthermore been noted that, even in the presence of first-order skews, second-order deviations from first-order predictions are often identical in complementary strands (Mrazek and Karlin, 1998). This is an indication that dinucleotide and trinucleotide selection and exclusion affect complementary strands independently of the context and symmetrically.

Likewise, at still higher orders, regulatory motifs on the order of 5 to 10

Table 4. Symmetry levels for a pool of coding regions of *S. cerevisiae*. The pool size is approximately 9 Mbp. Columns correspond to the order N at which symmetry is measured. The first row (labeled 0) shows the S^1 symmetry levels of Nmer distributions ($N = 1$ to 9). Successive rows show residual symmetry levels (S^1 computed on differences between predicted and observed distributions) after factoring out distributions of order $M = 1$ to 8

	1	2	3	4	5	6	7	8	9
0	0.94	0.93	0.90	0.88	0.86	0.84	0.81	0.79	0.74
1		0.85	0.64	0.65	0.62	0.60	0.57	0.55	0.51
2			0.32	0.45	0.45	0.44	0.43	0.41	0.39
3				0.63	0.56	0.52	0.48	0.44	0.40
4					0.42	0.38	0.38	0.36	0.33
5						0.31	0.35	0.34	0.32
6							0.39	0.35	0.32
7								0.31	0.29
8									0.28

bp are likely to be selected against except in proximity to genes they regulate. The investigation of the genomes of *S. cerevisiae* and *D. melanogaster* reveals that some Nmers are symmetrically over-represented at specific locations in upstream regions, and otherwise symmetrically under-represented (to be reported elsewhere). Computer simulations (not shown) confirm that the symmetric exclusion or selection of a subset of oligonucleotides promotes symmetry at all orders.

Table 5. Symmetry levels for a pool of non-coding regions upstream of genes in *S. cerevisiae*. The pool size is approximately 2.5 Mbp. See legend of Table 4

	1	2	3	4	5	6	7	8	9
0	0.99	0.99	0.98	0.98	0.97	0.96	0.94	0.91	0.84
1		0.91	0.90	0.90	0.88	0.86	0.82	0.75	0.65
2			0.88	0.89	0.87	0.84	0.79	0.72	0.62
3				0.93	0.88	0.83	0.76	0.66	0.57
4					0.79	0.74	0.67	0.59	0.52
5						0.67	0.61	0.55	0.49
6							0.57	0.51	0.47
7								0.47	0.45
8									0.43

Finally, our separate analysis of pooled sequences shows that eukaryotic coding regions display an asymptotic S^1 symmetry level of approximately 95% at the first order, while non-coding regions can approach perfect symmetry as the size of the pool increases. The asymptotic limit of coding regions, which is presumably linked to protein-coding constraints and transcription-related mutation biases, is for instance 94.0% in *S. cerevisiae* and 95.7% in *C. elegans*. While evidencing an intrinsic moderate first-order asymmetry of coding regions, these results also show that large-scale features can be substantially or almost perfectly symmetric. Therefore, an even strand repartition of coding and non-coding regions is not necessary to achieve substantial symmetry at the first order. Furthermore, as exemplified in Tables 4 and 5, plain and residual high-order symmetry drops sharply in coding regions as N increases, which contrasts with the asymptotic convergence towards perfect symmetry at all orders in large complete chromosomes and in the upstream regions of yeast. As an exception to this general rule, distinct convexities are often observed in residual symmetry profiles when order 3 and, to a lesser extent, order 6 are factored out. Presumably, asymmetric constraints linked to protein coding are then removed. The substantial first-order symmetry of coding regions is therefore achieved despite relatively asymmetric high-order constraints, and is thus likely to result from low-order mechanisms.

In short, three lines of evidence clearly indicate that low-order mechanisms contribute to strand symmetry: (a) the length scale at which substantial symmetry is often reached; (b) the evidence for symmetric constraints of orders 1 to 10; (c) the symmetry levels of coding and upstream regions.

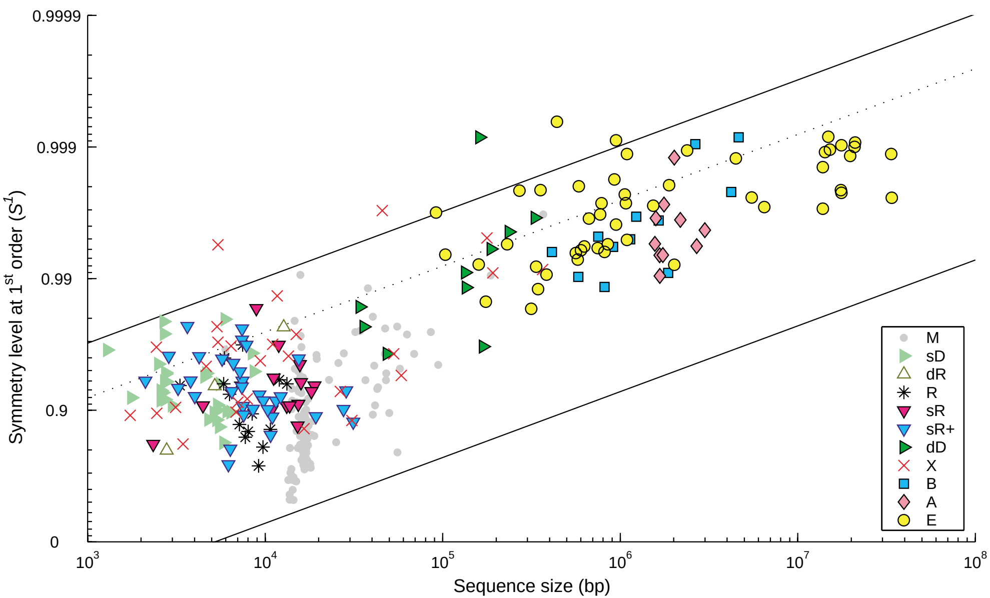

Fig. 1. Symmetry levels (S^1) measured on the base composition of 396 chromosomes or nucleic acid sequences of increasing length. Markers differentiate the following taxa or categories: viruses, single-stranded DNA (sD), double-stranded RNA (dR), retroids (R), single-stranded RNA- (sR-), single-stranded RNA+ (sR+), double-stranded DNA (dD); bacteria (B); archaea (A); eukaryota (E); extra-chromosomal elements (X); mitochondria (M). The upper line represents S^1 expectation for DNA sequences generated with a first-order, uniform (A,C,G,T) Markov model. The dotted line is 3 standard deviations away from the expectation. The corresponding upper bound exceeds the limits of the y-axis. The lowest line represents the average symmetry level that the model would yield for heptanucleotides. The lines were drawn after analytical approximations (**Methods**). They are in good agreement with simulation results. Any non-uniform first-order model yields lines that are parallel and close to the plotted ones.

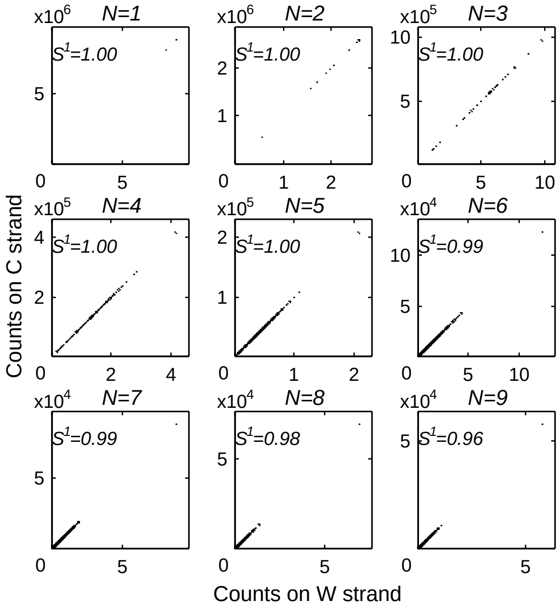

Fig. 2. Counts on direct strand versus counts on reverse-complementary strand of human chromosome 22, for oligonucleotides of length N=1 to 9. S^1 symmetry levels are reported on each plot. S^C symmetry levels amount to 1.00 at all orders.

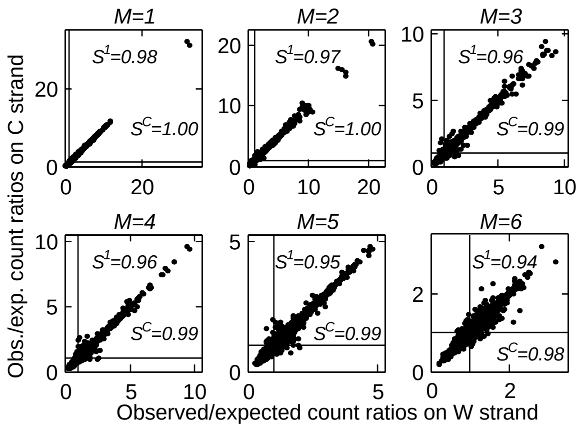

Fig. 3. Residual symmetry at order $N = 7$ in human chromosome 22. Deviations of observed heptanucleotide counts from those predicted by lower-order distributions are measured as ratios. Deviations observed on the direct strand are plotted against those observed on the reverse-complementary strand. Successive plots correspond to increasing predicting orders M. In the absence of high-order constraints, all points should lie around position $(1, 1)$, where two perpendicular lines intersect. The tight alignment of point along the diagonal translates in high S^1 and S^C levels, and shows that high-order phenomena contribute to the heptanucleotide symmetry that Fig. 2 ($N = 7$) illustrates.

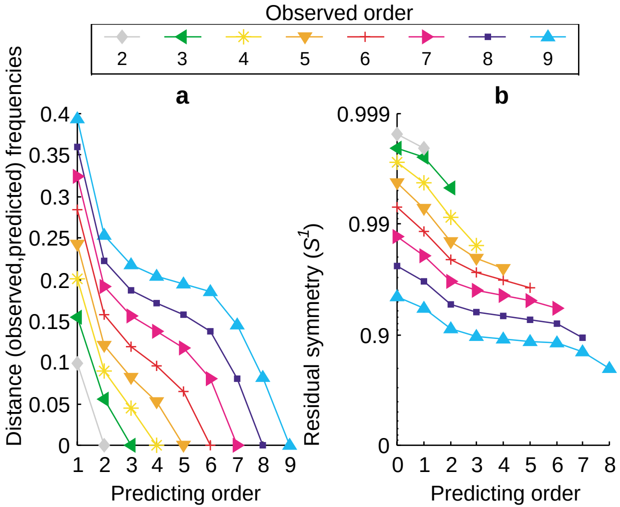

Fig. 4. Distance and residual symmetry profiles at orders 2 to 9 in human chromosome 22. (**a**) Distance. Each line represents the distance ($1 - S^1$) between the observed oligonucleotide frequencies at a given order N, and those predicted by the distributions of smaller oligonucleotides of length $M = 1$ to $N-1$. (**b**) Residual symmetry. Each line represents residual symmetry levels (S^1) for a given observed order N. The first point (labeled 0) represents the symmetry level of the distribution itself. Successive points (labeled 1 to 8) represent the symmetry level of its deviations (measured as differences) from the frequencies predicted by distributions of lower order $M = 1$ to $N - 1$.

Discussion

Through the methods we have developed, we have established that strand symmetry is a general emerging property of large poly-nucleotide molecules, that genuine high-order constraints promote symmetry at multiple orders, and that commonly accepted simple explanations of reverse-complement parities are inadequate. Pervasive strand symmetries must be considered as the compound effect of a wide spectrum of mechanisms that operate at multiple orders, leave their symmetric imprint at multiple length scales, and tend to shape complementary strands as well as functionally similar but non-contiguous regions.

Any selective pressure favoring intra-strand self-complementarity in relation to DNA, rRNA, tRNA or mRNA secondary structure, increases symmetry. Beyond such direct effects, it is important to realize that, provided it is blind to strandedness, any mechanism that alters double-stranded DNA or RNA sequences tends to promote reverse-complement symmetry. Consequently, symmetry does not necessarily represent a direct constraint or add a selective advantage *per se*. At the first-order, single-point insertions, deletions or substitutions thus generally result in approximately symmetric sequences. The symmetric selection and exclusion of Nmers 2 to 10 nucleotide long, notably in relation to their structural or binding properties, are also likely to significantly contribute to symmetry. At even higher orders, the even distribution of large-scale features among strands, which recombination events can promote through strand inversions, tends to cancel at a large scale the typically moderate asymmetries that can develop locally, notably in relation to the functions born by each strand, and to the corresponding mutation, repair, or signal sequence distribution biases.

Some of our results support the view that such very high-order mechanisms play a significant role. For instance, the equi-partition of genes between strands can cancel out in eukaryotes the intrinsic 5% asymmetry of coding regions at the first order. Counts of genes and the measurement of their base-pair coverage show that an even repartition is indeed often achieved in eukaryotes. Furthermore, base composition symmetry can display correlations with gene symmetry that are substantial and stronger than with sequence length. In the case of the 16 nuclear chromosomes of yeast, for instance, the correlation amounts to 0.76, and to 0.86 when including the mitochondrial chromosome. Chromosome 3, which contains mating loci, is the least symmetric (1.69% asymmetry at the first order) and also displays a particularly high gene asymmetry (15%). Likewise, an even repartition of large-scale features contributes to the high global symmetry levels observed in some mitochondria, viruses and prokaryotes, where first-order skews develop over long stretches. Although in a few cases skews might in principle be compensated by a mutual cancellation of transcription- and replication-related biases (Frank and Lobry, 1999), an even strand repartition of genes and leading/lagging replication regions is in general necessary to remove asymmetries. Within our data set, significant skews are found, for instance, in *E. coli* (Blattner *et al.*, 1997; Lobry, 1996a), *B. subtilis* (Lobry, 1996a) and adenoviruses (Grigoriev, 1999), for which global first-order S^1 symmetry nonetheless reaches 99.9%, 99.8% and 98%, respectively. In non-eukaryotic genomes, we again find substantial correlations between base-composition symmetry levels and the evenness of gene repartition among strands. For instance, the correlation coefficient is 0.5 in mitochondrial genomes, in which local asymmetries correlate both to gene orientation and to replication direction.

Duplication events followed by strand inversion promote symmetry at a large scale, and at all orders up to the size of the duplicated feature. Gene duplication, the multiplication of repetitive elements (among which palindromic, inherently symmetric sequences are over-represented (Cox and Mirkin, 1997)) and chromosomal inversions are thus likely contributors. The insertion of transposons and retroviruses is also a significant potential source of symmetry in higher eukaryotes, where they are relatively evenly distributed among strands and can represent more than 50% of the genome.

Some viruses (DNA and double-stranded RNA viruses) display high levels of symmetry with respect to their length, and single-stranded viruses and retroids are not strikingly less symmetric than double-stranded higher organisms once length effects are factored out (Table 3). While their integration to the host genome or their reliance on its replication machinery might explain their relative symmetry, viruses represent good candidates to assess the effect of putative selective pressures for self-complementarity within a strand.

Interestingly, we found that over- or under-representation profiles and residual symmetry profiles such as those exemplified in Fig. 4 are very similar

Abstract
Introduction
Materials and methods
Results
Discussion
Acknowledgements
References
Appendix
GO BACK
CLOSE FILE

for all chromosomes of a given organism. In addition, the similarity of over- or under-representation profiles and that of Nmer distributions are generally higher between complementary strands than between strands belonging to different chromosomes. This suggest that: (a) genome signatures—specific dinucleotide and trinucleotide over- and under-representation patterns that reflect phylogeny, and are measured in pooled complementary sequences—extend to high orders; (b) over sufficiently long windows, genome signatures are a property of single strands rather than pooled complementary strands; (c) the same spectrum of mechanisms that yield homogeneous Nmer distributions across chromosomes also shape single strands and promote symmetry.

The methods we have developed do not quantify the relative contribution of different mechanisms and orders to reverse-complement symmetry. Such an assessment is a matter for future investigation.

Acknowledgements

We thank the UCI Computational Genomics Group for helpful discussions. The work of PB is in part supported by a Laurel Wilkening Faculty Innovation award at UCI and a Sun Microsystems award. The work of SH was supported by a grant from the Chao Cancer Foundation.

References

Baldi,P. and Brunak,S. (2001) *Bioinformatics: the machine learning approach*, Second edition. MIT Press, Cambridge, MA.

Bell,S.J. and Forsdyke,D.R. (1999a) Accounting units in DNA. *J. Theor. Biol.*, **197**, 51–61.

Bell,S.J. and Forsdyke,D.R. (1999b) Deviations from Chargaff's second parity rule correlate with direction of transcription. *J. Theor. Biol.*, **197**, 63–76.

Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2000) GenBank. *Nucleic Acids Res.*, **28**, 15–18. MEDLINE Abstract

Blattner,F.R., Plunkett,G., Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F., Gregor,J., Davis,N.W., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J., Mau,B. and Shao,Y. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474. MEDLINE Abstract

Burge,C., Campbell,A.M. and Karlin,S. (1992) Over- and under-representation of short oligonucleotides in DNA sequences. *Proc. Natl Acad. Sci. USA*, **89**, 1358–1362. MEDLINE Abstract

Burland,V., Plunkett,Jr,G., Daniels,D.L. and Blattner,F.R. (1993) DNA sequence and analysis of 136 kilobases of the *Escherichia coli* genome: organizational symmetry around the origin of replication. *Genomics*, **16**, 551–561. MEDLINE Abstract

Campbell,A., Mrazek,J. and Karlin,S. (1999) Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proc. Natl Acad. Sci. USA*, **96**, 9184–9189. MEDLINE Abstract

Chargaff,E. (1951) Structure and function of nucleic acids as cell constituents. *Fed. Proc.*, **10**, 654–659.

Chargaff,E. (1979) How genetics got a chemical education. *Ann. NY Acad. Sci.*, **325**, 344–360. MEDLINE Abstract

Cox,R. and Mirkin,S.M. (1997) Characteristic enrichment of DNA repeats in different genomes. *Proc. Natl Acad. Sci. USA*, **94**, 5237–5242. MEDLINE Abstract

Fickett,J.W., Torney,D.C. and Wolf,D.R. (1992) Base compositional structure of genomes. *Genomics*, **13**, 1056–1064. MEDLINE Abstract

Forsdyke,D.R. (1995a) Relative roles of primary sequence and (G + C)% in determining the hierarchy of frequencies of complementary trinucleotide pairs in DNAs of different species. *J. Mol. Evol.*, **41**, 573–581.

Forsdyke,D.R. (1995b) A stem-loop 'kissing' model for the initiation of recombination and the origin of introns. *Mol. Biol. Evol.*, **12**, 949–958.

Forsdyke,D.R. and Mortimer,J.R. (2000) Chargaff's legacy. *Gene*, **261**, 127–137. MEDLINE Abstract

Frank,A.C. and Lobry,J.R. (1999) Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene*, **238**, 65–77. MEDLINE Abstract

Grigoriev,A. (1999) Strand-specific compositional asymmetries in double-stranded DNA viruses. *Virus Res.*, **60**, 1–19. MEDLINE Abstract

Hampson,S., Baldi,P., Kibler,D. and Sandmeyer,S.B. (2000) Analysis of yeast's ORF upstream regions by parallel processing, microarrays, and computational methods. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 190–201. MEDLINE Abstract

Kano-Sueoka,T., Lobry,J.R. and Sueoka,N. (1999) Intra-strand biases in bacteriophage T4 genome. *Gene*, **238**, 59–64. MEDLINE Abstract

Karkas,J.D., Rudner,R. and Chargaff,E. (1968) Separation of *B. subtilis* DNA into complementary strands. II. Template functions and composition as determined by transcription with RNA polymerase. *Proc. Natl Acad. Sci. USA*, **60**, 915–920. MEDLINE Abstract

Karkas,J.D., Rudner,R. and Chargaff,E. (1970) Template properties of complementary fractions of denatured microbial deoxyribonucleic acids. *Proc. Natl Acad. Sci. USA*, **65**, 1049–1056. MEDLINE Abstract

Karlin,S. and Burge,C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.*, **11**, 283–290. MEDLINE Abstract

Karlin,S. and Mrazek,J. (1997) Compositional differences within and between eukaryotic genomes. *Proc. Natl Acad. Sci. USA*, **94**, 10227–10232. MEDLINE Abstract

Lin,H.J. and Chargaff,E. (1967) On the denaturation of deoxyribonucleic acid. II. Effects of concentration. *Biochim. Biophys. Acta.*, **145**, 398–409. MEDLINE Abstract

GO BACK
CLOSE FILE

Lobry,J.R. (1995) Properties of a general model of DNA evolution under no-strand-bias conditions. *J. Mol. Evol.*, **40**, 326–330. MEDLINE Abstract

Lobry,J.R. (1996a) Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.*, **13**, 660–665.

Lobry,J.R. (1996b) Origin of replication of *Mycoplasma genitalium*. *Science*, **272**, 745–746.

Lobry,J.R. and Lobry,C. (1999) Evolution of DNA base composition under no-strand-bias conditions when the substitution rates are not constant. *Mol. Biol. Evol.*, **16**, 719–723. MEDLINE Abstract

Magasanik,B. and Chargaff,E. (1989) Studies on the structure of ribonucleic acids. 1951 [classical article]. *Biochim Biophys Acta*, 17–33. MEDLINE Abstract

Mrazek,J. and Karlin,S. (1998) Strand compositional asymmetry in bacterial and large viral genomes. *Proc. Natl Acad. Sci. USA*, **95**, 3720–3725. MEDLINE Abstract

Perna,N.T. and Kocher,T.D. (1995) Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes. *J. Mol. Evol.*, **41**, 353–358. MEDLINE Abstract

Prabhu,V.V. (1993) Symmetry observations in long nucleotide sequences. *Nucleic Acids Res*, **21**, 2797–2800. MEDLINE Abstract

Qi,D. and Cuticchia,A.J. (2001) Compositional symmetries in complete genomes. *Bioinformatics*, **17**, 557–559. MEDLINE Abstract

Abstract
Introduction
Materials and methods
Results
Discussion
Acknowledgements
References
Appendix
GO BACK
CLOSE FILE

Rudner,R., Karkas,J.D. and Chargaff,E. (1968a) Separation of *B. subtilis* DNA into complementary strands, I. Biological properties. *Proc. Natl Acad. Sci. USA*, **60**, 630–635.

Rudner,R., Karkas,J.D. and Chargaff,E. (1968b) Separation of *B. subtilis* DNA into complementary strands. 3. direct analysis. *Proc. Natl Acad. Sci. USA*, **60**, 921–922.

Rudner,R., Karkas,J.D. and Chargaff,E. (1969) Separation of microbial deoxyribonucleic acids into complementary strands. *Proc. Natl Acad. Sci. USA*, **63**, 152–159. MEDLINE Abstract

Schuler,G.D., Epstein,J.A., Ohkawa,H. and Kans,J.A. (1996) Entrez: molecular biology database and retrieval system. *Methods Enzymol.*, **266**, 141–162. MEDLINE Abstract

Sueoka,N. (1995) Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *J. Mol. Evol.*, **40**, 318–325. MEDLINE Abstract

Tillier,E.R. and Collins,R.A. (2000) The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. *J. Mol. Evol.*, **50**, 249–257. MEDLINE Abstract

Wu,C.I. and Maeda,N. (1987) Inequality in mutation rates of the two strands of DNA. *Nature*, **327**, 169–170. MEDLINE Abstract

Appendix

Symmetry of restrictions

Consider a symmetric Markov model of order N which induces a symmetric distribution on Nmers so that for any Nmer $P(X_1 \dots X_N) = P(\bar{X}_N \dots \bar{X}_1)$. Consider now the lower-order restriction of this distribution to Mmers, with $M < N$. The probability distribution induced on the space of Mmers is also symmetric. To see this, the probability of any Mmer according to the higher-order distribution is given by:

$$P(X_1 \dots X_M) = \sum_{Y_1 \dots Y_{N-M}} P(X_1 \dots X_M Y_1 \dots Y_{N-M}). \tag{3}$$

For the reverse complement, we have:

$$\begin{aligned} P(\bar{X}_M \dots \bar{X}_1) &= \sum_{Z_1 \dots Z_{N-M}} P(Z_1 \dots Z_{N-M} \bar{X}_M \dots \bar{X}_1) \\ &= \sum_{Y_1 \dots Y_{N-M}} P(\bar{Y}_{N-M} \dots \bar{Y}_1 \bar{X}_M \dots \bar{X}_1) \\ &= P(X_1 \dots X_M), \end{aligned} \tag{4}$$

the last equality resulting from the symmetry of the distribution of order N.

Expectation and variance of S^1

If X is a $\mathcal{N}(0, \sigma^2)$ normally distributed random variable, with mean 0 and variance σ^2, then it is easy to check that the expectation and the variance of the random variable $|X|$ are:

$$E(|X|) = \sqrt{\frac{2}{\pi}}\sigma \quad \text{and} \quad \mathrm{Var}(|X|) = \left(\frac{\pi - 2}{\pi}\right)\sigma^2. \tag{5}$$

Consider now Nmer frequency values f_i and f_i'. Assume, for approximation purposes, that the difference of frequencies $f_i - f_i'$ is normally distributed with mean 0 and variance σ_d^2. Then, by linearity of the expectation and using (Equation 5), the expected value of the symmetry index $S^1 = 1 - \sum_i |f_i - f_i'|/2$ can be approximated by:

$$E(S^1) \approx 1 - \frac{4^N}{\sqrt{2\pi}}\sigma_d. \tag{6}$$

For simplicity, we can look at the case of a sequence with total length L and a uniform Nmer distribution where f_i is approximately normal with mean $1/4^N$ and variance $\sigma^2 = (1 - 1/4^N)/(L4^N)$. It is reasonable then to approximate the distribution of $f_i - f_i'$ by $\mathcal{N}(0, \sigma_d^2 = 2\sigma^2)$. Substituting in (Equation 6) gives in this case:

$$E(S^1) \approx 1 - \frac{\sqrt{4^N - 1}}{\sqrt{\pi L}}. \tag{7}$$

Notice that $E(S^1)$ decreases with N. Clearly as $L \to \infty$, $E(S^1) \to 1$. For large orders N, we also have $E(S^1) \approx 1 - 2^N/\sqrt{\pi L}$.

For the variance, if we ignore the small covariance, we get: $\mathrm{Var}(S^1) \approx 4^N \mathrm{Var}(|f_i - f_i'|)/4$. Using (Equation 5), we get:

$$\mathrm{Var}(S^1) \approx \frac{1}{2}\frac{\pi - 2}{\pi}\frac{4^N - 1}{4^N L} \approx \frac{1}{2}\frac{\pi - 2}{\pi L}. \tag{8}$$

The variance goes to 0 with the length like $1/L$.

S^1 *decreases monotonically with* N

Consider S^1_N and S^1_{N+1}, the symmetry measures of a sequence at order N and $N + 1$. For any Nmer $X_1 \ldots X_N$ and ignoring boundary effects, overlapping counts give immediately: $P(X_1 \ldots X_N) = \sum_Y P(X_1 \ldots X_N Y)$. Using this and the triangle inequality in the formula for S^1 shows that the numerator in Equation 1 in general increases substantially (in some trivial cases it may remain constant) when going from order N to order $N + 1$. The denominator, on the other hand, remains constant and equals 2 when computed on distributions. Thus the value of S^1 decreases with the order N. It is easily verified that when a distribution of order M is perfectly symmetric ($S^1 = 1$) or perfectly asymmetric ($S^1 = 0$), then its factorial extension to any higher order N yields the same S^1 value.

Residual S^1 value after removal of all symmetry

A standard calculation shows that:

$$\int_{-\infty}^{+\infty} |x| \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} dx = \frac{\mu}{\sigma}\left[-1 + 2F\left(\frac{\mu}{\sigma}\right)\right] + \frac{2\sigma}{\sqrt{2\pi}} e^{-\mu^2/2\sigma^2}, \tag{9}$$

where F is the cumulative distribution of the normalized Gaussian. Now suppose that we are approximating the distribution f of order N by the distribution g, and f' by g'. In a typical case, g and g' are the factorial distributions resulting from lower-order Markov models of complementary strands. When we are trying to factor out the effect of g on f, we look at how close $f-g$ is typically to 0 or how close f/g is to 1. To a first approximation we can model these deviations $f-g$ or f/g using a Gaussian $\mathcal{N}(\mu, \sigma^2)$. When g is a very good approximation to f, in particular when g is the Markov model of order N, then $\mu = 0$ (resp. $\mu = 1$) for the difference (resp. for the ratio). When g comes from a strictly-lower order, then the residual mean may not necessarily be 0. In all cases, $(f-g)-(f'-g')$ can then be approximated by a Gaussian $\mathcal{N}(0, 2\sigma^2)$. Using Equations (5) and (9), we have the rough approximations:

$$S^1 \approx 1 - \frac{\sqrt{\frac{2}{\pi}}\sqrt{2}\sigma}{2\left[\frac{\mu}{\sigma}\left[-1+2F\left(\frac{\mu}{\sigma}\right)\right] + \frac{2\sigma}{\sqrt{2\pi}} e^{-\mu^2/2\sigma^2}\right]}. \tag{10}$$

When $\mu = 0$, this gives $S^1 \approx (\sqrt{2} - 1)/\sqrt{2} \approx .293$. Thus the residual S^1 symmetry one measures at order N on a DNA sequence generated using a symmetric Markov model of order $M < N$, after factoring out the distribution of order M, is approximately .293 at any reasonable length L. Essentially the same value is obtained from (Equation 10) when the ratio is used instead, with $\mu = 1$ provided σ is reasonably small so that $F(\mu/\sigma) \approx 1$ and $\exp(-\mu^2/2\sigma^2) \approx 0$. The residual S^1 value captures a random background level of overlap in the deviations from the factorial extensions, measured on complementary strands.