

Bias of Purine Stretches in Sequenced Chromosomes

David Ussery*, Dikeos Mario Soumpasis, Søren Brunak
Hans Henrik Stærfeldt, Peder Worning, and Anders Krogh

Center for Biological Sequence Analysis
Department of Biotechnology
Building 208
The Technical University of Denmark
DK-2800 Kgs. Lyngby, Denmark

January 14, 2002

*to whom correspondence should be addressed. Tel: (+45) 45 25 24 88; Fax: (+45) 45 93 15 85; email: dave@cbs.dtu.dk

ABSTRACT. We examined more than 700 DNA sequences (full length chromosomes and plasmids) for stretches of purines (R) or pyrimidines (Y) and alternating YR stretches; such regions will likely adopt structures which are different from the canonical B-form. Since one turn of the DNA helix is roughly 10 bp, we measured the fraction of each genome which contains purine (or pyrimidine) tracts of lengths of 10 bp or longer (hereafter referred to as "purine tracts"), as well as stretches of alternating pyrimidines/purine ("pyr/pur tracts") of the same length. Using this criteria, a random sequence would be expected to contain about 0.2% percent of purine tracts and also about 0.2% of the alternating pyr/pur tracts. Both motifs were found to be over-represented. In the vast majority of cases, there are more purine tracts than would be expected from a random sequence, with an average of 3.5%, significantly larger than the expectation value. The fraction of the chromosomes containing pyr/pur tracts was also over-represented, although to a lesser extent, with an average of 0.8%. One of the most surprising findings is a clear difference in the length distributions of the regions studied between prokaryotes and eukaryotes. Whereas short-range correlations can explain the length distributions in prokaryotes, in eukaryotes there is an abundance of long stretches of purines or alternating purine/pyrimidine tracts, which cannot be explained in this way; these sequences are likely to play an important role in eukaryotic chromosome organisation.

A-DNA / Z-DNA / DNA structures / complete genomes / Bacterial chromosomes / Eukaryotic chromosomes / Viral chromosomes

1 Introduction

There are three main families of DNA helices: A-DNA, B-DNA, and Z-DNA. In solution, for most DNA sequences the helical structure is a mixture of the A- and B- conformations. Certain sequences – in particular GC-rich alternating pyrimidine/purine (YR) stretches can form the left-handed Z-DNA conformation within the context of flanking right-handed genomic DNA (Sinden & Kochel, 1987). Depending on the sequence, environmental conditions and biomolecular interactions, genomic DNA can adopt various conformations of A, B, and Z structural families (Saenger *et al.*, 1986; Sinden, 1994; Foloppe & Jr., 1999). Although it is generally agreed that the B-form is predominant under physiological conditions, different structures may locally exist in different regions of genomic DNA. These structures can be utilised in subtle ways to influence gene expression, gene regulation, and chromatin structure, in conjunction with DNA-protein binding equilibria and in response to local environmental changes (Sinden, 1994).

Since it has been established experimentally that purine tracts and alternating pyr/pur tracts can form helical conformations other than the canonical B-form, we have examined sequenced chromosomes for such motifs as a first step in estimating the amount DNA in non-B-form conformations.

1.1 Purine stretches and A-DNA

Stretches of G's (or C's) longer than four nucleotides prefer the A-conformation, provided that the energetics of the A/B junctions are sufficiently low (that is, with favourable flanking sequences) (Ng *et al.*, 2000). In contrast, stretches of A's (T's) do not convert to the A-form but adopt a distinct right-handed form (B') stabilised by a network of water molecules, the so-called spine of hydration observed in x-ray studies (Kopka *et al.*, 1983) and computed in theoretical work (Garcia *et al.*, 1996). Stretches of G's (or C's) are over-represented in eukaryotic chromosomes (Behe, 1998).

A-form helices are common for DNA-RNA hybrids, as well as for double stranded RNA. Regions of DNA which code for stable RNA genes, such as rRNAs and tRNAs, have more of a tendency to form an A-like structure, not necessarily from the point of view of stability of the DNA, but from the fact that a nucleotide sequence which forms a stable RNA double stranded molecule might also form a more stable A-DNA conformation (Antony *et al.*, 1999). Another

common biological occurrence of sequences which can readily form A-DNA is in viral Long Terminal Repeats (LTRs) (Mujeeb *et al.*, 1993). These regions often contain purine stretches which favour the A-DNA conformation (Suzuki *et al.*, 1996). It is likely that these regions are involved in recombination, perhaps through a triple-stranded DNA intermediate. The A-DNA helix can readily accommodate a third strand of DNA into the major groove (Sekharudu *et al.*, 1993).

1.2 Alternating pur/pyr stretches and Z-DNA

In general, pyr/pyr stretches are thermodynamically less stable than other DNA sequence motifs. Runs of (TA)_n will readily melt, and (CG)_n repeats can form Z-DNA. Clusters of (CG)_n, of at least 6 nucleotides or longer, correspond with experimentally determined regions of Z-DNA (Konopka *et al.*, 1985). One of the first crystal structures of DNA was the left-handed Z-DNA for CGCGCG (Wang *et al.*, 1979). Sequences which can form Z-DNA are essentially not found in *E. coli* chromosomes, and yet they are over-represented in the chromosomes of many eukaryotes. A notable example of this is the CpG islands, which could potentially form Z-DNA, especially when methylated. In a complicated scenario, a protein which is responsible for mRNA editing is activated upon binding to left-handed Z-DNA upstream of a gene (Schade *et al.*, 1999). There have been many other biological roles postulated for Z-DNA, include acting a transcriptional enhancer (Banerjee & Grunberger, 1986); the left-handed Z-conformation might furthermore be involved in terminal differentiation (Gagna *et al.*, 1999). Finally, Z-DNA could also be involved in recombination (Majewski & Ott, 2000).

Structural, physicochemical and theoretical studies have elucidated many aspects of left-handed Z-DNA and its stability relative to the ubiquitous B-form (Herbert & Rich, 1999; Jovin *et al.*, 1987; Soumpasis, 1984). It is well established that alternating CG stretches (4 bp in length or longer) adopt Z-forms under favorable conditions such as high concentrations or monovalent salts, low concentrations of multivalent ions, chemical modifications such as C methylation, and topological stress induced by supercoiling. Similar behaviour is also observed (to a somewhat lesser degree) with CA tracts, also depending on the nature of the sequences flanking the RY stretch. Various computer programmes have been developed to search for potential Z-DNA regions within sequences (Ho *et al.*, 1986), although extension of programmes which takes into account coopera-

tivity of the sequence environment to human chromosomes of hundreds of millions of bp in length is difficult.

AT stretches can adopt conformations of the alternating B-type but do not seem to convert to Z-DNA under conditions favorable for the formation of left-handed DNA, probably due to differential hydration effects stabilising the right-handed conformation in TA tracts. However, it should be noted that TA tracts can melt much more readily, and also can form cruciforms, under the right conditions (Darlow & Leach, 1995).

1.3 Measurement of levels of purine and pur/pyr stretches in sequenced chromosomes

In view of the knowledge accumulated to date it is clear that the occurrence of local structures alternative to the B-form along genomic DNA cannot be accurately predicted from knowledge of the sequence without substantial incorporation of suitable parameterisations of physicochemical data, as well as theoretical and algorithmic advances likely to emerge in the near future. However, in view of the explosion of sequences currently produced by worldwide genomic projects, it is clearly interesting and appropriate to perform a large scale analysis of the data using available computational tools, with the aim to localise regions where alternative structures may occur, evaluate their statistical significance and compare their distributions in various organisms. In a first approximation, this can be done using the most elementary sequence signatures involved – namely sufficiently long (e.g., 10 bp or longer) alternating (YR) or homogeneous (R or Y) stretches, bearing in mind that these sequence patterns are necessary but not sufficient for local occurrence of alternative structures in genomic DNA due to the sequence context and physicochemical dependencies involved.

In this work, we shall measure the occurrence of purine (or pyrimidine) tracts of at least 10 bp in length (henceforth called "purine tracts"), and compare this with alternating pyrimidine/purine tracts, also of at least 10 bp in length ("pyr/pur tracts"). In view of the level of analysis presently possible, more detailed structural classifications are not justified. We have examined the occurrence of purine and pyr/pur stretches in more than 700 publicly available fully sequenced chromosomes or plasmids, and present the results in 2 different forms: first, as a simple ratio of the

Kingdom	no. chromosomes/plasmids	total no. bp
Archaea	20	26,409,849 bp
Bacteria	165	208,304,570 bp
Proctista	18	18,053,080 bp
Fungi	23	36,117,519 bp
Plants	7	47,623,657 bp
Animals	36	2,979,841,298 bp
Viruses	491	12,279,171 bp

Table 1: Summary table of the chromosome and plasmid sequences downloaded from GenBank. The number of individual DNA sequences examined includes plasmids and organelles for sequenced genomes. The "total number of bp" reflects the count of A,T,G, and Cs within the sequence, and does not include "n" or other ambiguous bases.

percentage of the chromosome which contains these stretches, and second, as a "DNA atlas" plot, where these regions can be localised and visualised within the context of the whole chromosome.

2 Methods

2.1 The data sets

We downloaded all the sequenced genomes from GenBank, including the plasmids and organelles associated with each sequenced genome. This resulted in a total of 764 sequenced chromosomes from Prokaryotic and Eukaryotic organisms, as well as double-stranded DNA viruses. The totals can be seen in Table 1. Links to the individual chromosome sequences and references can be found on our "Genome Atlas" web page ¹.

2.2 Calculation of A-DNA and Z-DNA fractions in chromosomes

Homopurine (or homopyrimidine) tracts of at least 10 bp in length were determined from the GenBank files from sequenced chromosomes, using a simple PERL script to search for regular expressions, and write a single line for each base in the sequence – if the particular sequence was part of a purine stretch of given length, it was given a score of 1, otherwise it was given a score of 0. The fraction of A-DNA is simply the sum of 1's for this file, divided by the total length of the chromosome. The file can also be used as input for the Atlas plots to generate a visual localisation of the stretches, in terms of the whole annotated chromosome.

The expected values were estimated by calculating the probability of a purine (or pur/pyr)

¹<http://www.cbs.dtu.dk/services/GenomeAtlas/A.DNA/>

tract occurring by chance in a random sequence. This is one-half to the power of the length of the stretch times the probability that the stretch is interrupted, $\left(\frac{1}{2}\right)^{n+2}$. Thus, the probability that a given nucleotide is part of n purines or pyrimidines in a row is $n \cdot \left(\frac{1}{2}\right)^{n+1}$, where we multiplied by two because the stretch could occur on either strand. The probability that a nucleotide is part of a purine or pyrimidine stretch length n or longer, $P(n)$, is then:

$$P(n) = \sum_{i=n}^{\infty} i \cdot \left(\frac{1}{2}\right)^{i+1} = 1 - \sum_{i=1}^{n-1} i \cdot \left(\frac{1}{2}\right)^{i+1} = (n+1) \left(\frac{1}{2}\right)^n. \quad (1)$$

The same expression is valid for an alternating purine-pyrimidine tract of length n bp.

The above calculation ignores correlations between nucleotides. A Markov chain of order m takes local short-range correlations into account. According to such a model, the probability of $n > m$ purines ($R_1 \dots R_n$) flanked by pyrimidines (Y) and preceded by $m - 1$ arbitrary bases ($N_1 \dots N_{m-1}$), is:

$$P(R_1 \dots R_n Y | N_1 \dots N_{m-1} Y) = \text{constant} \cdot P(R | R \dots R)^n, \quad (2)$$

where the last probability is the Markov chain probability of a purine occurring after m purines. Similar expressions holds for pur/pyr tracts.

Thus, if a genome is approximately Markovian of a reasonable order², the length distribution of such stretches will be linear in a logarithmic plot. A deviation from linearity would suggest a functional preference for such regions.

2.3 GenomeAtlas plots

We have created special ‘‘DNA Atlas’’ plots (Pedersen *et al.*, 2000; Jensen *et al.*, 1999). to visualise purine stretches and pur/pyr regions throughout the genome. The ‘‘A-DNA Atlas’’ localises homopurine (R) n and homopyrimidine (Y) n tracts, whilst the ‘‘Z-DNA Atlas’’ visualises regions of alternating pyrimidine/purine (YR) n tracts.

²Here a ‘reasonable order’ means that the order is less than the lengths of interest, i.e., around 10. It would not make sense to think about Markov chains of much higher order, because the number of parameters would exceed the total number of base pairs in the genome (at order 20 the number of free parameters is approximately the size of the total human genome (3×10^9)).

3 Results and Discussion

3.1 Are purine and pur/pyr tracts a result of short-range correlations?

Examples of logarithmic length distributions are shown Figure 2 for two prokaryotic and two eukaryotic chromosomes. In both the bacterium *Escherichia coli* (Figure 2A), and the Archaea *Pyrococcus furiosus* (Figure 2B), the curves are close to linear over many decades suggesting that long purine or pur/pyr tracts can be explained by short-range correlations. The curves are quite close to the line obtained from the simple Bernulli model that assumes independent bases. These curves are representative for all prokaryotes.

In the case of eukaryotic chromosomes, the situation is dramatically different, as can be seen by the plots for the yeast genome (Figure 2C) and for human chromosome 1 (Figure 2D), where there are clear and significant overrepresentation of long purine and pyr/pur stretches. The deviation from the line clearly shows that human DNA is not well explained by a homogeneous model of local composition such as a Markov chain. The deviation from the line could be explained by the heterogeneity of human DNA. However, we have observed the same effect in other eukaryotes with much less heterogeneity (such as, for example, the yeast genome shown in Figure 2C), which suggests that the long tracts observed can not be explained by short-range correlations. It is likely that the abundance of long purine or pyr/pur tracts is a result of DNA structural preferences.

For human chromosome 1 we estimated a sixth order Markov chain and generated a random sequence of the same length as the chromosome. The logarithmic length distributions of purine and pyr/pur tracts are also shown in Figure 2D. They are linear as expected and agree well with the initial part of the distributions for short lengths. This is just to emphasize the deviation of the distributions from anything resulting from short-range correlations.

3.2 Bias in chromosome sequences towards purine stretches

Figure 1 shows the results for estimated levels of purine stretches and pur/pyr stretches in chromosomes. In all of the various Kingdoms examined, there are more of these sequences than would be expected from a simple model of nucleotide distribution. Furthermore, in all Kingdoms the occurrence of purine stretches is more abundant than for pyr/pur stretches. Bacteria and Viruses contain the least amount of these stretches, which might be partly reflective of coding constraints on their genomes.

3.3 Purine and pur/pyr stretches in prokaryotic chromosomes

3.3.1 Archaea

We examined the sequenced genomes of 13 Archaeal organisms, and the results are shown in Figure 3. With the single exception of *Halobacterium* species NRC1, all other Archaeal genomes contained much larger fractions of purine than pyr/pyr stretches. *Halobacterium* differs from the other 12 Archaea examined in that it lives in an extreme high-salt environment. One possible explanation for this difference might be that purine stretches are providing a structural role by adopting an A-DNA conformation under "normal conditions". Since A-DNA can readily form even for mixed-sequence DNA under high-salt conditions, the need for purine stretches to stabilise it under lower salt concentrations would not be necessary (Feig & Pettitt, 1999). The sequences of more strongly halophylic organisms are needed to further test this hypothesis. Many of the Archaeal genomes have quite high levels of purine stretches. For example, the three different species of *Pyrococcus* all have more than 5% A-DNA (about 25-fold larger than the expected value of about 0.2%). However, (again with the exception of the salt-loving *Halobacterium*) nearly all the Archaeal genomes contain much lower values of pyr/pyr stretches than found in chromosomes from other organisms. In fact, as can be seen in Figure 1, Archaea contain the less amount of pyr/pyr stretches than any of the other Kingdoms. However, Archaea is a very diverse group, and only a relatively few genomes have been sequenced, so it is perhaps too soon to draw strong conclusions about the relative amounts of purine and pyr/pyr stretches in this Kingdom.

3.3.2 Bacteria – Proteobacter

Figure 4 shows the observed purine and pyr/pyr frequencies for chromosomes from 24 different proteobacterial species. These species fall into five different taxonomic groups, designated alpha through epsilon. Currently, there is only one sequenced chromosome from the delta subdivision, but some trends can be seen in Figure 4 for the remaining four subdivisions.

The alpha subdivision includes nitrogen-fixing bacteria (members of the rhizobium group) as well as several other types of organisms, such as intracellular parasites. The first eight chromosomes all contain more pyr/pyr than purine stretches, but the *Rickettsia* chromosomes contain the opposite ratio.

The beta subdivision contains three sequenced Neisserial genomes, as well as that of *Bor-*

detella pertussis. The three Neisserial genomes all have very similar frequencies of pyr/pur and purine tracts, but the *Bordetella* genome has a quite high fraction of pyr/pur tracts (more than 3%), with a comparatively small fraction of purine tracts.

The gamma subdivision contains many commonly known bacteria, such as *Escherichia coli*. It is interesting to note that the large *E. coli* and *Salmonella* plasmids have an abundance of purine stretches. The *Pasteurella*, *Pseudomonas*, *Vibrio*, and *Xylella* subgroups each also have their own patterns, as can be seen in Figure 4.

The epsilon subdivision of proteobacteria (e.g., the three far-right chromosomes in Figure 4) includes the pathogens *Helicobacter pylori* and *Campylobacter jejuni*. All three chromosomes show a high amount of purine stretches (more than 2.5%, or more than the average of any of the other proteobacter subdivisions) as well as very low levels of pyr/pur stretches. These organisms are AT rich (*Helicobacter pylori* is 61% AT, *Campylobacter jejuni* is 69% AT), although there are other genomes within the proteobacter group with similar AT content, but different distributions of purine and pyr/pur stretches. For example, *Rickettsia prowazekii*, a member of the alpha subdivision has 70% AT content, and yet it has a different profile.

3.3.3 Bacteria – Firmicutes

Frequencies of purine and pyr/pur tracts for 27 firmicute chromosomes are shown in Figure 5. The chromosomes are sorted by AT content, with *C.diphtheria* having 27% AT, and *U. urealyticum* containing 74% AT. The five Actinobacteria genomes have less than 50% AT content, whilst the remaining bacteria (from the "Bacillus/Clostridium" group) have more than 50% AT content. All five of the Actinobacteria chromosomes contain larger fractions of pyr/pur stretches than purine stretches, and the opposite is true for the other Firmicute chromosomes. This is not likely to be due merely to a difference in AT content, since many other chromosomes with less than 50% AT content show different trends (for example, compare the profile in Figure 6 of the Green-sulfur bacterium *Chlorobium tepidum*, which has a 42% AT content).

3.3.4 Other Bacteria

The remaining bacterial chromosomes from organisms which are neither Proteobacter or Firmicutes are shown in Figure 6. Both the *Aquifex aeolicus* VF5 main chromosome and its plasmid

contain around 8% purine stretches, which is amongst the highest of any of the more than 700 sequences examined. Similarly, the genome of *Thermotoga maritima* MSB8 contains quite high levels of purine stretches and low amounts of RY stretches. Both of these organisms are thermophilic bacteria, believed to be related to ancient "primitive" bacteria. All eight of the *Chlamydia* and all 25 of the Spirochaetales sequences (labelled in Figure 6) contain around 3% purine stretches and around 0.5% pyr/pur stretches.

3.4 Purine and pyr/pur tracts are Over-represented in Eukaryotes

Figure 2D shows a clear bias in the distribution of purine and pyr/pur stretches in human chromosome 1. All the eukaryotic chromosomes we examined have significantly more purine and pyr/pur stretches than expected. This is true even for *C. elegans*, which is known to have a compact genome (for an animal), and, in all fairness, none of the chromosomes have been completely sequenced. There are still many gaps (represented by long stretches of N's) in the GenBank sequence, and it is possible that some of these contain pyr/pur stretches which might influence the total fraction. All of the animal and plant chromosomes available at the time of writing include large gaps which could well influence the numbers reported here. That is, it is likely that some of the repeats currently not included contain such motifs as YR tracts (for example, CpG islands are difficult to sequence).

However, several completely sequenced microbial chromosomes of about the same size as bacterial chromosomes are available. Figure 2 shows the levels of purine and pyr/pur stretches for chromosomes from the kingdom Protocista. This group contains more than a dozen chromosomes that have been completely sequenced, without any gaps. The levels of purine and pyr/pur stretches from the protozoan chromosomes are quite high – in the case of *Plasmodium falciparum* around 8% (that is, more than 40-fold larger than the expected value).

3.5 Localisation of purine and pyr/pur stretches to intergenic regions in protozoan chromosomes

Where are these RY stretches localised? Although *Leishmania major* does not contain many introns, there seems to be a strong localisation of pyr/pur stretches to intergenic regions. A "DNA Atlas" plot for chromosome 1 from *Leishmania major* (Myler *et al.*, 1999) is shown in Figure 7A. The top four lanes (A-D) indicate the density of simple tetrameric repeats (G4, A4, T4, and C4,

respectively), followed by the annotated coding regions. Lanes F and G plot strand preference for pyrimidine tracts, and lane H is the AT content of the sequence. Note that there is a clear localisation of purine stretches in intergenic regions of the chromosome. For comparison, the DNA Atlas for pyr/pur stretches of the same chromosome is shown in Figure 7B. Note that the purine stretches, whilst still localised mainly in intergenic regions, are quite distinct in their patterns from the purine stretches shown in Figure 7A. In the case of *Leishmania*, roughly half of the DNA is coding for proteins, and the other half is non-coding. Localisation of alternative DNA structures to the non-coding half makes it seem likely that the purine and pyr/pur regions are playing a structural role in the chromosome organisation.

3.6 Summary

We have found three items of interest by comparing the levels of purine tracts and pyr/pur stretches. First, there is a clear difference between the Markovian behaviour of prokaryotic DNA compared to eukaryotics. Second, in nearly all organisms examined, there is an over-representation of purine stretches of 10 bp in length or longer; this is particularly pronounced in the completely sequenced protozoan chromosomes. Finally, in microbial eukaryotic chromosomes, the purine and pyr/pur stretches can be localised mainly to non-coding regions of the chromosome.

Acknowledgements

This work was supported by a grant from the Danish National Research Foundation. The authors would like to thank Al Ivens from the Sanger Centre for helpful discussions about *Leishmania* chromosomes.

References

- Antony, T., Thomas, T., Shirahata, A. & Thomas, T. (1999). Selectivity of polyamines on the stability of rna-dna hybrids containing phosphodiester and phosphorothioate oligodeoxyribonucleotides. *Biochemistry*, **38**, 10775–10784.
- Banerjee, R. & Grunberger, D. (1986). Enhanced expression of the bacterial chloramphenicol acetyltransferase gene in mouse cells cotransfected with synthetic polynucleotides able to form Z-DNA. *Proc. Natl. Acad. Sci. USA*, **83**, 4988–4992.

- Behe, M. (1998). Tracts of adenosine and cytidine residues in the genomes of prokaryotes and eukaryotes. *DNA Seq.*, **8**, 375–383.
- Darlow, J. & Leach, D. (1995). The effects of trinucleotide repeats found in human inherited disorders on palindrome inviability in *Escherichia coli* suggest hairpin folding preferences in vivo. *Genetics*, **141**, 825–832.
- Feig, M. & Pettitt, B. (1999). Modeling high-resolution hydration patterns in correlation with dna sequence and conformation. *J. Mol. Biol.*, **286**, 1075–1095.
- Foloppe, N. & Jr., A. M. (1999). Intrinsic conformational properties of deoxyribonucleosides: implicated role for cytosine in the equilibrium among the a, b, and z forms of dna. *Biophys. J.*, **76**, 3206–3218.
- Gagna, C., Kuo, H. & Lambert, W. (1999). Terminal differentiation and left-handed Z-DNA: a review. *Cell Biol. Int.*, **23**, 1–5.
- Garcia, A., Hummer, G. & Soumpasis, D. (1996). Theoretical description of biomolecular hydration: Application to DNA. In Schoenborn, B. & Knott, A., (eds.) *Neutrons in Biology*. Plenum Press, New York, NY, pp. 299–308.
- Herbert, A. & Rich, A. (1999). Left-handed Z-DNA: structure and function. *Genetica*, **106**, 37–47.
- Ho, P., Ellison, M., Quigley, G. & Rich, A. (1986). A computer aided thermodynamic approach for predicting the formation of z-dna in naturally occurring sequences. *EMBO J.*, **5**, 2737–2744.
- Jensen, L., Friis, C. & Ussery, D. (1999). Three views of the *E. coli* genome. *Research in Microbiology*, **150**, 773–777.
- Jovin, T., Soumpasis, D. & McIntosh, L. (1987). The transition between B-DNA and Z-DNA. *Ann. Rev. Phys. Chem.*, **38**, 521–560.
- Konopka, A., Reiter, J., Jung, M., Zarling, D. & Jovin, T. (1985). Concordance of experimentally mapped or predicted Z-DNA sites with positions of selected alternating purine-pyrimidine tracts. *Nucl. Acids Res.*, **13**, 1683–1701.

- Kopka, M., Fratini, A., Drew, H. & Dickerson, R. (1983). Ordered water structure around a B-DNA dodecamer: a quantitative study. *J. Mol. Biol.*, **163**, 129–146.
- Majewski, J. & Ott, J. (2000). Gt repeats are associated with recombination on human chromosome 22. *Genome Res.*, **10**, 1108–1114.
- Mujeeb, A., Kerwin, S., Kenyon, G. & James, T. (1993). Solution structure of a conserved dna sequence from the hiv-1 genome: restrained molecular dynamics simulation with distance and torsion angle restraints derived from two-dimensional nmr spectra. *Biochemistry*, **32**, 13419–13431.
- Myler, P., Audleman, L., de Vos, T., Hixson, G., Kiser, P., Magness, C., Rickel, E., Sisk, E., Sunkin, S., Swartzell, S., Westlake, T., Bastein, P., Fu, G., Ivens, A. & Stuart, K. (1999). *Leishmania major* friedlin chromosome 1 has an unusual distribution of protein-coding genes. *Proc. Natl. Acad. Sci. USA*, **96**, 2902–2906.
- Ng, H., Kopka, M. & Dickerson, R. (2000). The structure of a stable intermediate in the a to b DNA helix transition. *Proc. Natl. Acad. Sci. USA*, **97**, 2035–2039.
- Pedersen, A., Jensen, L., Stærfeldt, H., Brunak, S. & Ussery, D. (2000). A DNA structural atlas of *E. coli*. *J. Mol. Biol.*, **299**, 907–930.
- Saenger, W., Hunter, W. & Kennard, O. (1986). DNA conformation is determined by economics in the hydration of phosphate groups. *Nature*, **324**, 385–388.
- Schade, M., Turner, C., Kuhne, R., Schmieder, P., Lowenhaupt, K., Herbert, A., Rich, A. & Oschkinat, H. (1999). The solution structure of the zalpha domain of the human RNA editing enzyme ADAR1 reveals a prepositioned binding surface for Z-DNA. *Proc. Natl. Acad. Sci. USA*, **96**, 12465–12470.
- Sekharudu, C., Yathindra, N. & Sundaralingam, M. (1993). Molecular dynamics investigations of DNA triple helical models: unique features of the watson-crick duplex. *J. Biomol. Struct. Dyn.*, **11**, 225–244.
- Sinden, R. & Kochel, T. (1987). Reduced 4,5',8-trimethylpsoralen cross-linking of left-handed Z-DNA stabilized by DNA supercoiling. *Biochemistry*, **26**, 1340–1350.

Sinden, R. R. (1994). *DNA Structure and Function*. Academic Press, New York.

Soumpasis, D. (1984). Statistical mechanics of the B-Z transition of DNA: Contribution of diffuse ionic interactions. *Proc. Natl. Acad. Sci. USA*, **81**, 5116–5120.

Suzuki, M., Yagi, N. & Finch, J. (1996). Role of base-backbone and base-base interactions in alternating DNA conformations. *FEBS Lett*, **29**, 148–152.

Wang, A., Quigley, G., Kolpak, F., Crawford, J., van Boom, J., van der Marel, G. & Rich, A. (1979). Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature*, **282**, 680–686.

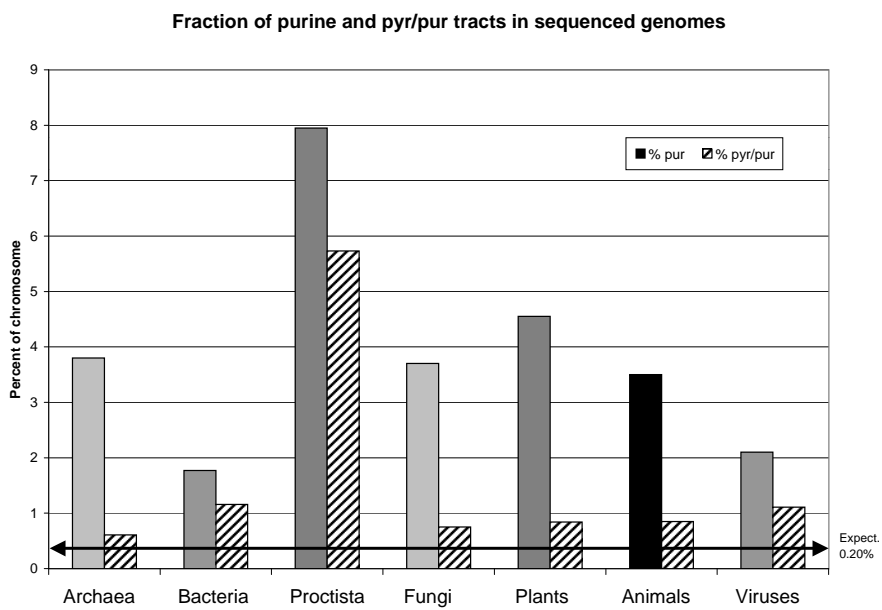
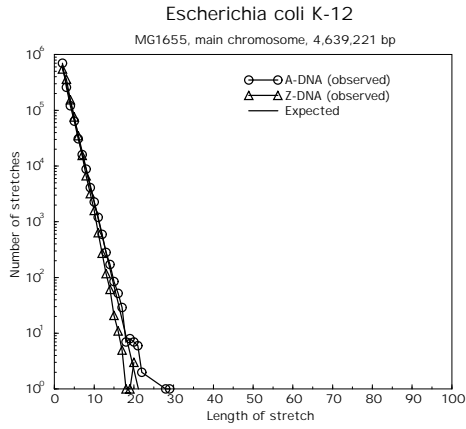
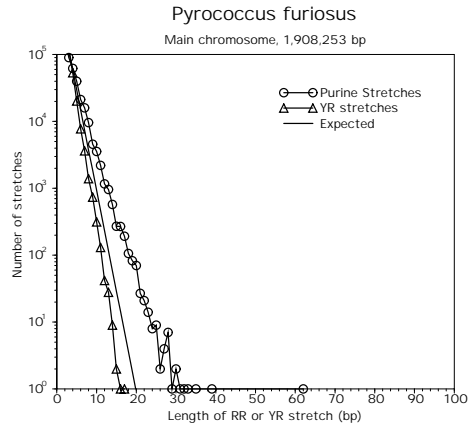


Figure 1: Observed frequencies of purine and pur/pyr stretches in various organisms. The percentage of "pur tracts" refer to the fraction of the chromosome which contains homopurine (or homopyrimidine) stretches of at least 10 bp in length, and the percentage of "pyr/pur tracts" is the fraction of the chromosome with alternating (YR) stretches of at least 10 bp in length.

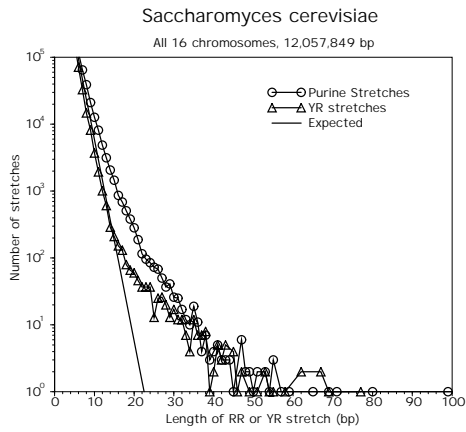
A.



B.



C.



D.

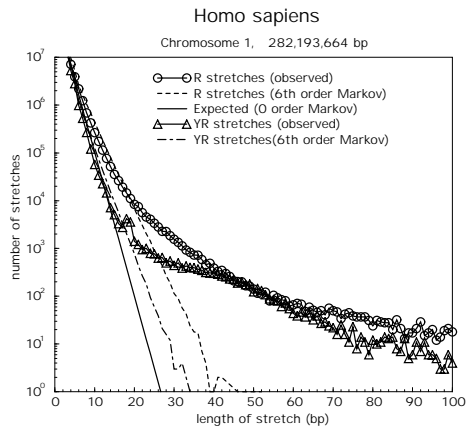


Figure 2: Observed vs. expected frequencies of purine and pyr/pur stretches in A. the *Escherichia coli* K-12 genome, B. the *Pyrococcus furiosus* genome, C. the *S. cerevisiae* genome, and D. human chromosome 1.

Purine and pur/pyr tracts in Archaeal Genomes

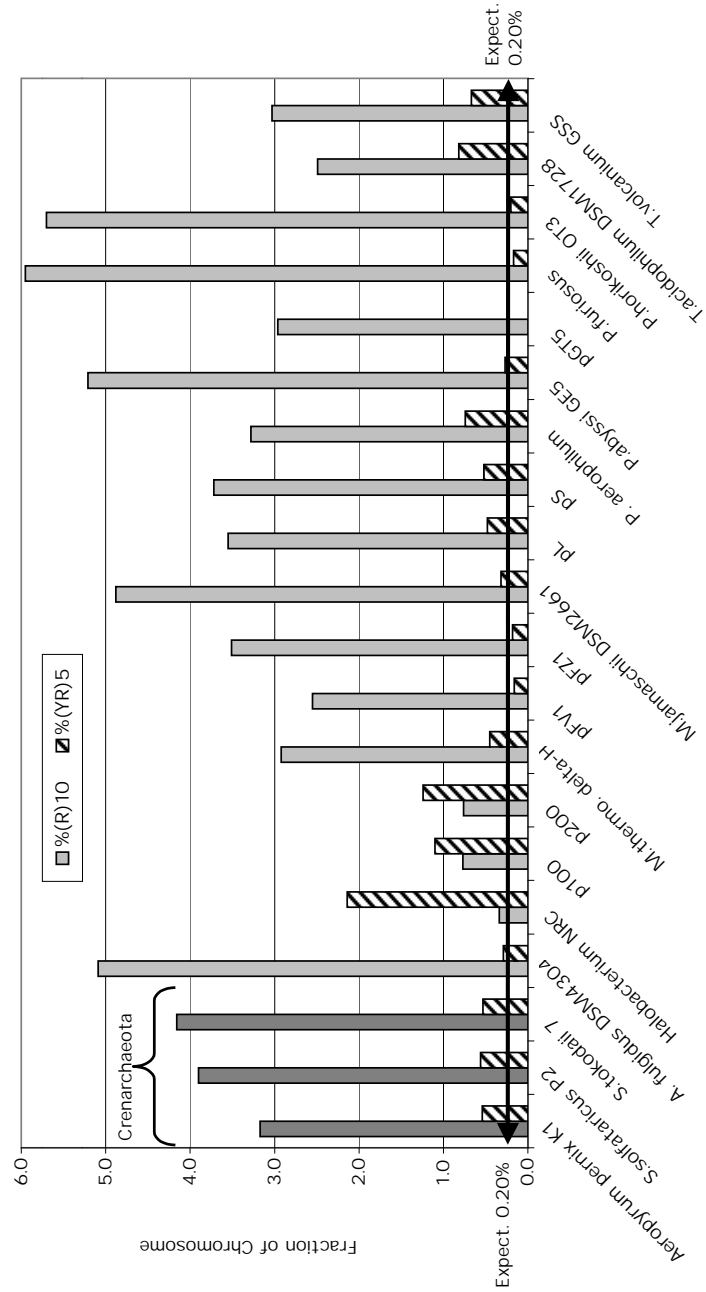


Figure 3: Fraction of purine and pur/pyr stretches in 13 sequenced Archaea genomes. The expected value is designated with an arrow. The first three chromosomes on the left are from Crenarchaeal genomes, as labelled; the remaining 10 genomes are Euarchaea.

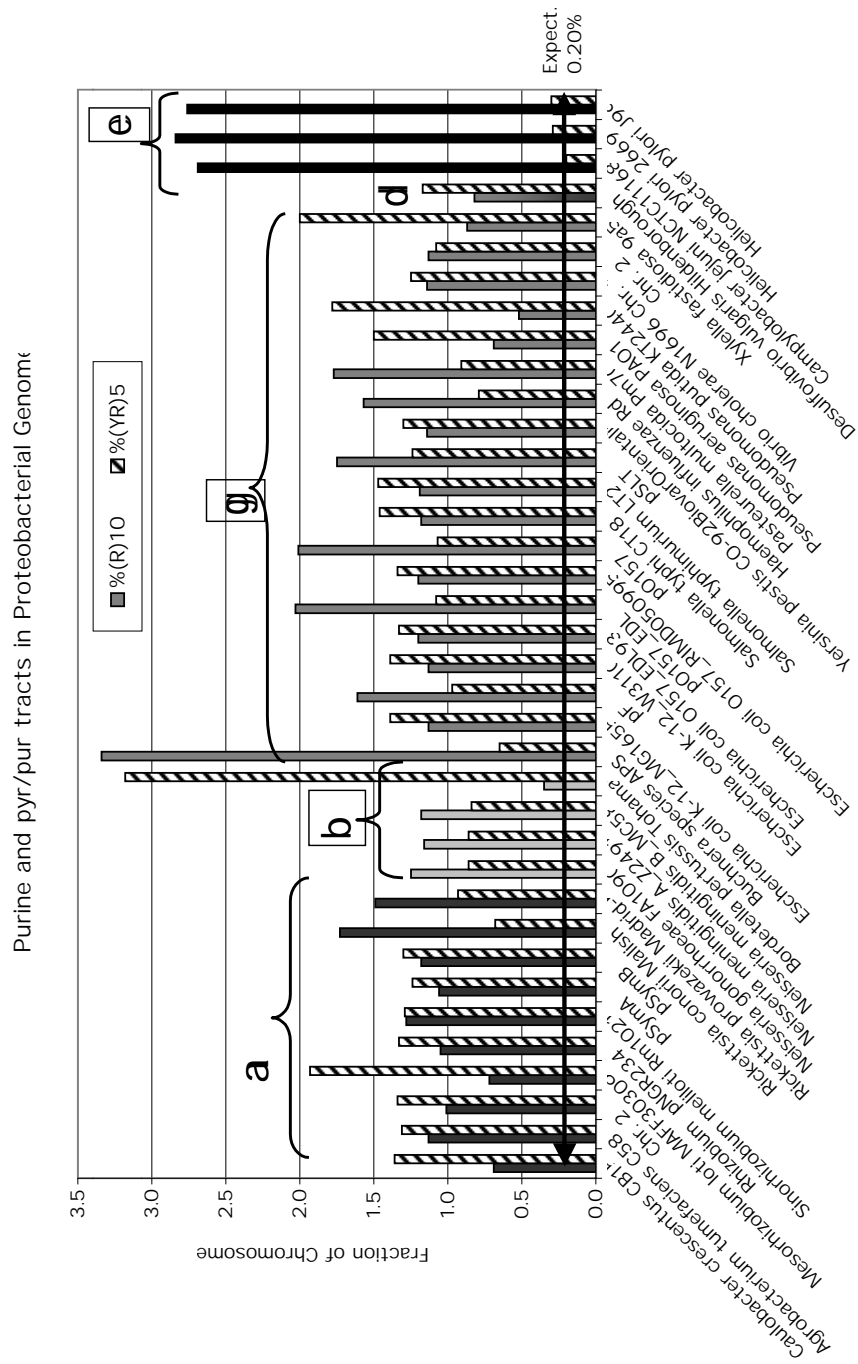
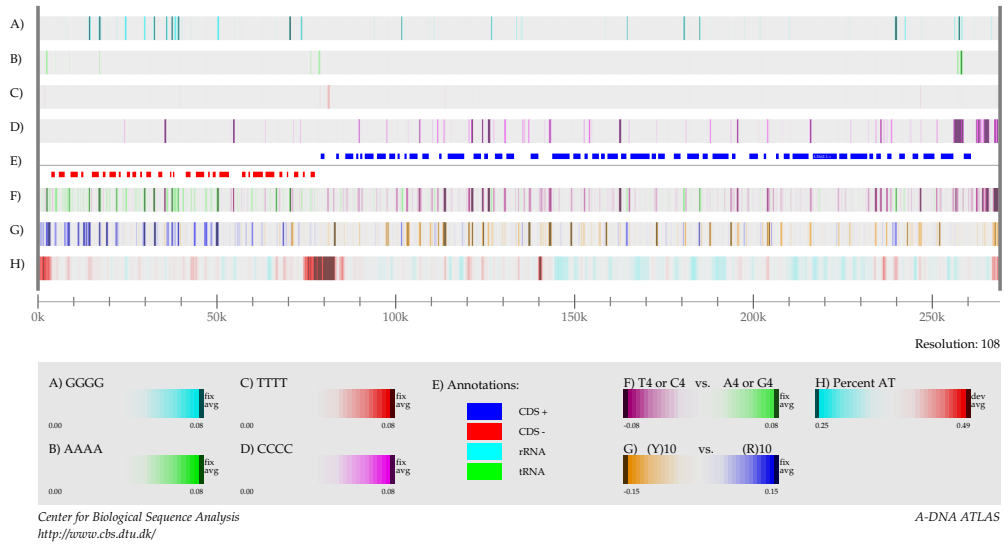


Figure 4: Fraction of purine and pyr/pur stretches in 24 proteobacterial species. The alpha, beta, gamma, delta, and epsilon subdivisions are as indicated.

A.

Leishmania major

Freidlin Chromosome I 268,984 bp



B.

Leishmania major

Freidlin Chromosome I 268,984 bp

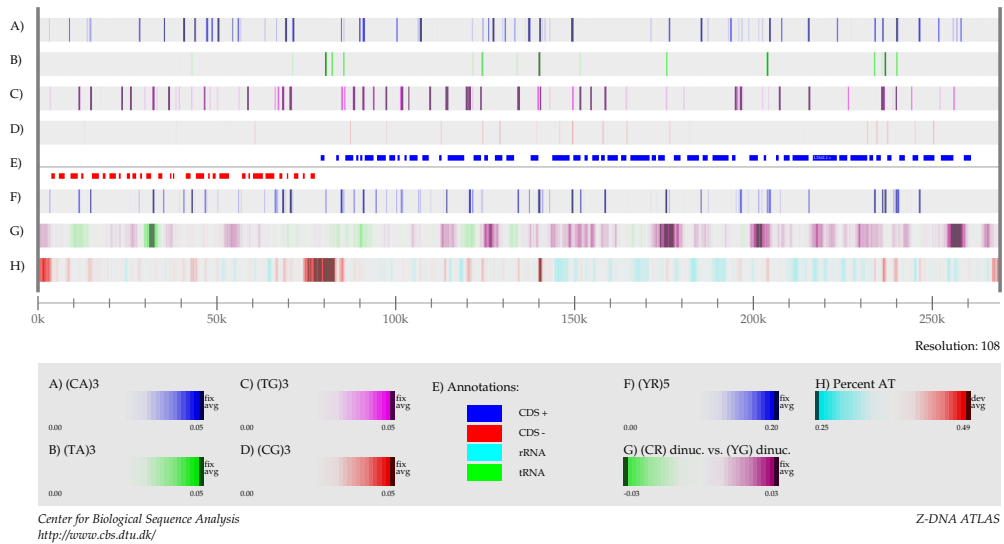


Figure 7: A. Localisation of purine stretches and B. localisation of pyr/pur stretches, within *Leishmania major* chromosome 1.