Nucleic Acids Research, 1993, Vol. 21, No. 18 4239–4245

Stability, structure and complexity of yeast chromosome III

Graham J.King
Breeding and Genetics Department, Horticulture Research International, Wellesbourne, Warwick
CV35 9EF, UK

Received June 7, 1993; Revised and Accepted July 28, 1993

ABSTRACT

The complete sequence of yeast chromosome III provides a model for studies relating DNA sequence and structure at different levels of organisation in eukaryotic chromosomes. DNA helical stability, intrinsic curvature and sequence complexity have been calculated for the complete chromosome. These features are compartmentalised at different levels of organisation. Compartmentalisation of thermal stability is observed from the level delineating coding/non-coding sequences, to higher levels of organisation which correspond to regions varying in G + C content. The three-dimensional path reveals a symmetrical structure for the chromosome, with a densely packed central region and more diffuse and linear sub-telomeric regions. This interspersion of regions of high and low curvature is reflected at lower levels of organisation. Complexity of *n*-tuplets (*n* = 1 to 6) also reveals compartmentalisation of the chromosome at different levels of organisation, in many cases corresponding to the structural features. DNA stability, conformation and complexity delineate telomeres, centromere, autonomous replication sequences (ARS), transposition hotspots, recombination hotspots and the mating-type loci.

INTRODUCTION

Yeast chromosome III is the first eukaryotic nuclear chromosome for which a complete linear DNA sequence has been determined [1]. As such, it provides a useful model for general studies relating DNA sequence and structure at different levels of organisation. Although small in size (315356 bp) it resembles chromosomes from multi-cellular eukaryotes both in structure and mechanisms of transcription, replication, recombination and segregation.

Since the publication of the total sequence, a number of studies have addressed the distribution of DNA sequences along the chromosome [2,3,4]. Sharp & Lloyd found that against a background of 35% G+C base composition, there were two regions in which G+C values rose above 50% in coding regions, but not in adjacent intergenic regions [3]. They proposed that observed constraints on base usage in different regions of the chromosome may be caused by replication timing or recombination frequency.

Increased G+C content is generally correlated with higher DNA helical stability. However, the helix-coil transition (denaturation) is dependent not only on overall base composition, but on base-stacking and long-range interactions. The opening of the DNA helix during *in vitro* denaturation reflects thermodynamically constrained processes such as transcription, replication and recombination. Denaturation occurs in a discrete and cooperative manner over segments of 30bp to several hundred bases [5]. The process may be accurately modelled using thermodynamic parameters and is usually plotted as a 'stability map', which due to the cooperative nature of the process appears greatly different from the local map of G+C content. Gene coding sequences tend to display a homostabilising propensity, whereby individual base pairs contribute to a region of uniform stability [6,7,8]. Such stability maps also differ in kind from measures based solely on nearest-neighbour dinucleotide stability values for short DNA duplexes [9,10] which do not take into account long-range interactions found in indefinitely long DNAs at high concentration.

A further constraint on DNA sequence composition in eukaryotic chromosomes is the requirement for the double helix to be packaged around nucleosomes and thence into higher order structures. Ordered packaging is accomplished by the action of chromosomal proteins, notably the histones, acting synergistically on the intrinsic curvature and flexibility conferred by the sequence of DNA bases. A number of models, including the wedge model [11], have been proposed to account for the intrinsic curvature and flexibility of the DNA helix, which are based on nearest neighbour interactions. Other models exist for prediction of histone octomer binding sites based either on vectorial addition of roll, or roll and tilt angles, or probability functions [12].

In addition, the DNA sequence is required to code for polypeptides. In small chromosomes one might expect the information-carrying properties of the DNA sequence to be maximised within the framework required for maintenance of essential chromosome structures and functions. However, Karlin and co-workers have demonstrated [4] a relative abundance of local and global repeats in yeast chromosome III, highlighting five genes containing close or tandem repeats, an anomalous distribution of delta (δ) elements, a significantly even distribution of autonomous replication sequences (ARS), a relative increase in frequency of T runs and AT iterations downstream of genes and A runs upstream of genes, and two regions of complex repetitive sequences and anomalous DNA composition. This suggests that a degree of order, or loss of complexity, is required for maintenance of the chromosome.

Information theory is the mathematical study of information rate, channel capacity, noise, and other factors affecting information transmission. The theory was originally developed by Shannon [13] for telecommunications, and has since been used to formalise the information carrying capacity of DNA [14,15,16]. The information carrying capacity of a DNA sequence has been defined in terms of 'entropy', or complexity. Sibbald and coworkers [17] refined previous approaches by regarding a maximally complex sequence (of infinite length) as having a C_{MAX} of log 4 = 2 bits/base. Such a sequence is regarded as having no obvious pattern or order, and thus there is no way accurately to guess the next base. Order or pattern can be found at various levels. It is useful to be able to partition this order into increasing levels, from nucleotides through dinculeotides etc., in order to determine at which levels and to what degree the sequence differs from having equally frequent and independently distributed *n*-tuplets. Such measures may then be related to the occurrence of gene-coding and control sequences and those required for chromosomal integrity and structure.

Understanding the relationship between information carrying capacity and structural constraints for a small chromosome such as yeast chromosome III may assist in analysing DNA sequences now being accumulated for larger chromosomes of humans or agricultural species, where coding sequences occur less frequently.

METHODS

The complete sequence of *Saccharomyces ceriviseae* chromosome III (GenBank accession X59720), and sequence feature tables were extracted from the GenBank database release using the PIR Experimental Query System (XQS) software. Open reading frames and coding orientation were checked against [1].

Calculation of%GC base composition was carried out using the program GC-CALC using an overlapping scanning window of 100 bp. Theoretical thermal melting profiles were calculated using a modification of the MELT87 [18] program kindly supplied by L.Lerman and W.Fripp (MIT). The algorithm for the probability of helicity of each base pair was taken from [19], using a value of 1.8 in the loop entropy exponent, with no correction for loop stiffness. The enthalpy values for the stability of the nearest-neighbour base-pair doublets were taken from [20] based on 0.02 M Na^+. The dissociation constant was calculated according to [21]. A temperature resolution of 0.01 °C was used and the temperature contour for 50% helicity plotted as a conventional 'melt-map'.

Calculation of curvature maps and three-dimensional DNA path was carried out using the 'CURVATUR' program [22,23] kindly provided by E.Shpigelman & E.Trifonov (Weizmann Institute). The parameters chosen for the curvature map were a window of 22bp, with no smoothing option. Data were averaged over 100bp for subsequent plotting. The coordinates for the three dimensional path of DNA were combined with the thermal stability to produce a conformational plot of the chromosome. A 'ribbon' was plotted orthogonal to the direction of curvature, with the thickness a decreasing function of stability. Thus unstable regions appear as a thicker region, whilst more stable regions appear as a fine strand. The plot was created as a Postscript file in 'Mathematica'.

Calculation of sequence complexity was carried out using the program 'COMPLEX1' written by the author in Pascal, following the algorithm described in [17].

The complexity at the nucleotide level was calculated as:

$$C_1 = -\sum_{i=A}^{T} f_i \log_2 f_i \qquad \text{bits/base,}$$

where the summation is over all 4 bases and f_i is the frequency of the *i*th base. When all $f_i = 0.25$ then $C_1 = 2$, since there is no order at the nucleotide level.

At the doublet level, complexity is measured as:

$$C_2 = -\sum_{i=A}^{T}\sum_{j=A}^{T} f_i f_j \log_2 (f_i f_j) - (-\sum_{k=AA}^{TT} f_k \log_2 f_k)$$

where *i* and *j* are the frequencies of single bases and f_k is the frequency of doublets. The *k* summation is over all 16 doublet frequencies. In general, for any *n* the complexity of the actual *n*-tuplets (the *k* summations above) is subtracted from the complexity of the expected *n*-tuplets. The expected *n*-tuplets are the product of the $n-1$ tuplet frequency corresponding to the first $n-1$ bases, divided by the frequency of the central $n-2$ tuplet common to both $n-1$ tuplets.

Thus at the hexanucleotide level:

$$C_6 = -\sum_{i=Awxyz}^{Twxyz}\ \sum_{j=wxyzA}^{wxyzT} f_i f_j / f_{wxyz} \log_2 (f_i f_j / f_{wxyz}) - (-\sum_{k=AAAAAA}^{TTTTTT} f_k \log_2 f_k)$$

A three-stage process was used for these calculations. The first involved calculating the frequencies of all *k* possible *n*-tuplets ($n=1$ to 6; $k= 4^n$) throughout the chromosome. These results were stored in a table for use in subsequent calculations. The complexity values for all *k* possible *n*-tuplets were then calculated and stored. Finally the complexity values for each *n*-tuplet were plotted for each base position along the chromosome.

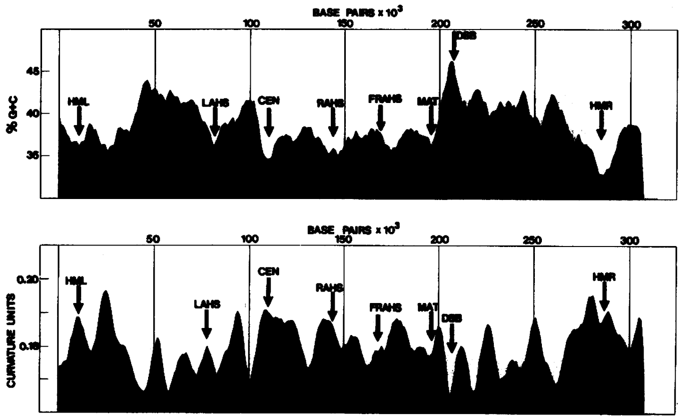

Figure 1. a. Distribution of G+C content and thermal stability along chromosome III. Percentage G+C values were initially calculated over a 100bp running window, and the values averaged over 2000bp for plotting. The profile coincides with the plot of thermal stability (see Figure 2) when values are averaged over 2000bp. **b.** Curvature map of chromosome III, calculated using the 'wedge' model [22,23]. The curvature units are an integration of wedge angle, helical twist and roll angles. One unit corresponds to the curvature required for one complete turn around a nucleosome. Features marked are: **HML**: *HML* silent mating type locus; **LAHS**: left arm transposition hotspot; **CEN**: centromere; **RAHS**: right arm transposition hotpsot; **FRAHS**: far right arm transposition hotspot; **MAT**: *MAT* mating type locus; **DSB**: double strand break site and recombination hotspot [24]; **HMR**: *HMR* silent mating type locus.

Data were plotted using programs written in Turbo Pascal or from Lotus 1-2-3.

RESULTS

Base composition

G+C content varies along the chromosome at different levels of organisation. Small-scale fluctuations covering several hundred to several thousand base pairs occur within gene-coding regions. Larger fluctuations are present in many non-coding sequences, with G+C values varying by up to 26%G+C over a 100bp stretch. Longer range variation is also observed (Figure 1a), which partitions the chromosome into several regions of distinct G+C content. The regions of highest G+C content are clustered in the two chromosome arms, consistent with the findings of Sharp & Lloyd [3]. In general the chromosome appears to be divided into regions, 25kbp to 30kbp in length, of discrete G+C content, flanked by shorter regions of lower G+C content. At a yet larger scale there are regions in both arms of about 75kbp which are clearly delineated as being of higher G+C content. These are separated by a 75kbp region of lower G+C content, which stretches from the centromere into the right arm. Coding sequences possess a higher and more uniform base composition than adjacent non-coding regions, although when the chromosome is regarded as a continuous stretch an almost random distribution is found, with a modal value of 37.5% G+C.

DNA stability

DNA helical stability calculated as predicted denaturation (Figure 2), displays a more structured pattern of variation along the chromosome than G+C content. Cooperatively melting regions occur throughout the sequence, with relatively uniform stability values over stretches of approximately 300bp to 2500bp. In general the more uniform regions coincide with coding sequences. There is a clear association between coding/non-coding boundaries and fluctuations in thermal stability of the helix. In addition, the non-coding regions are generally of lower and more variable stability than adjacent coding regions. The relationship between DNA helical stability and G+C base-composition is apparent only when averaged over at least 2000bp (Figure 1a), where the plots coincide. At the higher resolution (Figure 2), it is apparent that the gradual changes in helical stability along the chromosome are achieved step-wise by the coding regions, punctuated by relative instability in the intergenic regions.

Although in general the changes in helical stability appear to delineate coding regions, there are a few coding sequences which contain regions of high stability. These may compensate for longer low stability regions in the vicinity. The distribution of stability jumps is regular throughout the chromosome, with an average of 5500bp between large (2°C−5°C) stability changes.

DNA intrinsic curvature

The intrinsic curvature calculated using the 'wedge' model displays considerable local variation when plotted at high resolution as a 'curvature map' (not shown). No systematic pattern of variation emerges at this level. However, when averaged over longer sequences (up to 2000bp) a clearer pattern of regional variation emerges (Figure 1b). The chromosome is divided up into discrete regions, about 10kbp long, of increased curvature flanked by shorter regions of relatively 'straight' DNA. There does not appear to be any continuous gradient of curvature throughout the chromosome, although the central region possesses longer stretches of highly curved DNA. The predicted three dimensional path followed by the chromosomal DNA is shown in Figure 3. The pattern of interspersion of curved and relatively

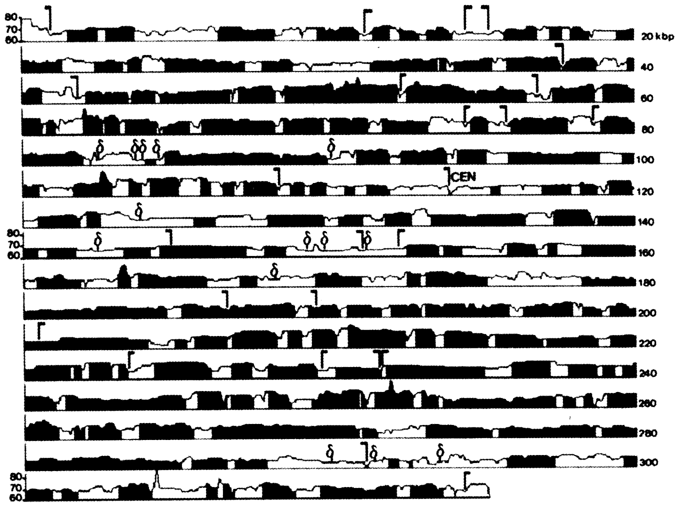

Figure 2. High-resolution map of thermal stability along chromosome III calculated using the MELT87 program [18]. The temperature contour for 50% helicity is plotted, with the y-axis in°C. Coding regions are shaded (dark shading Watson strand open reading frame (ORF), lighter shading Crick strand ORF). Delta elements are shown by δ; ARS consensus sequences by —.

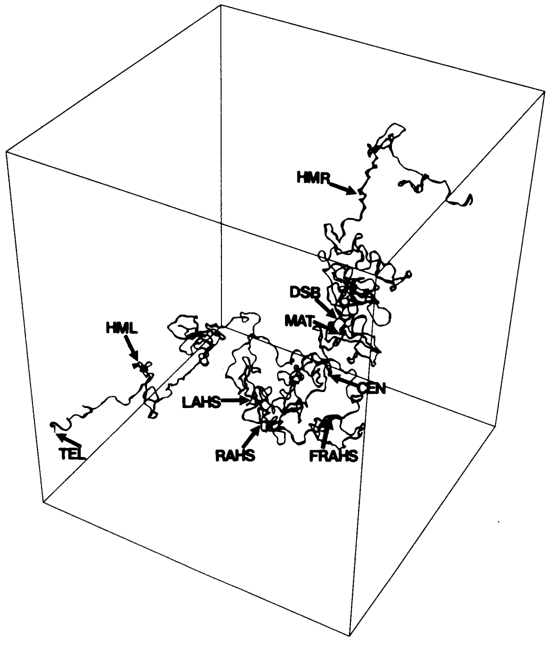

Figure 3. Conformational plot of three-dimensional path followed by 315356bp of the complete sequence of yeast chromosome III. The three dimensional coordinates were calculated using the wedge model [22,23]. The thickness of the DNA ribbon is plotted as a decreasing function of stability, calculated as in Figure 2.

straight DNA is observed throughout the chromosome. The central region is clearly more compact as a result of the higher intrinsic curvature. The sub-telomeric regions are noticeably less curved and thus less compact.

DNA sequence complexity and information content

Sequence complexity was calculated along the chromosomal sequence using a measure of information content based on the algorithms of Sibbald and coworkers [17] (Figure 4). The complexity accounted for by nucleotides, di-, tri-, tetra-, penta- and hexa-nucleotides can be seen to vary along the chromosome. At the nucleotide level there is clear compartmentalisation, with a central region of relatively uniform high information content corresponding to the 100kbp proximal to the centromere in the right arm. This is clearly delineated by short regions of relatively low complexity. There appears to be some indirect relationship between nucleotide complexity and curvature (compare Figures 1b and 4a), with a similar hierarchy of compartmentalisation.

At the dinucleotide level a compartmentalised distribution is also evident, although this does not coincide with the map of nucleotide complexity. At the trinucleotide level, shorter regions of about 30kbp are observed. The peak at 220kbp corresponds to the relative order at the nucleotide level. Similar subdivisions are also observed in the tetranucleotide and pentanucleotide level. At high resolution a number of low complexity hexanucleotide sequences are evident dispersed throughout the chromosome, which probably correspond to repeat motifs such as are found at the telomere.

Chromosomal structural features

The telomeres are characterised by high helical stability and low curvature. At the left telomere the CCCACACACCACA repeats contribute to a straight region which extends for 360bp (Figure 5a). The sequence data are not available for the last 300bp of the right telomere. The subtelomeric regions, extending for 12000bp, possess a low packing ratio compared to the central regions of the chromosome.

The centromere is clearly delineated as a region of low helical stability and high curvature (Figure 5b). The centromere region is flanked by regions of higher stability, and relatively low curvature, forming a loop type structure. In the context of the whole chromosome the centromere is positioned between two densely packed regions.

Chromosomal processes

Replication. There are 24 ARS consensus sequences (WTTTAYRTTTW) in the chromosome. Four of these (at 57051, 152338, 232099 & 291368bp) are in the midst of high repetitive and anomalous DNA structure [4]. Figure 2 demonstrates that in fact most ARS consensus sequences in the chromosome form islands of low helical stability and low local curvature, and are often flanked by regions of higher stability. Only a small number of the consensus sequences present are expected to be active sites of replication origin. The observed variation in the genomic environment probably contributes to the intrinsic ease of unwinding of the active origins. The DNA path flanking a number of ARS consensus sequences is characterised by a U-bend type pattern, where the ARS consensus element lies in the flat bottom of the U.

Recombination. Two regions have been highlighted as recombination hotspots in chromosome III. One has been mapped to the region around 211100 bp and coincides with a double-strand break (DSB) site [25]. The DSB is located at a peak of helical stability in the right arm, which also appears as a local kink in the DNA path, which otherwise is following a smooth contour as one arm of a longer loop (Figure 5d). This recombination hotspot, together with the centromere, delineates a central region of relatively high nucleotide complexity. It is also located at a peak of trinucleotide complexity, a phenomenon which may be related to the atypical codon usage as a result of the increased G+C content in the region. The other recombination hotspot is in the 12kbp region proximal to the left telomere, although this has not yet been finely mapped. This region is relatively uncurved, and is anchored by the stable telomere sequence.

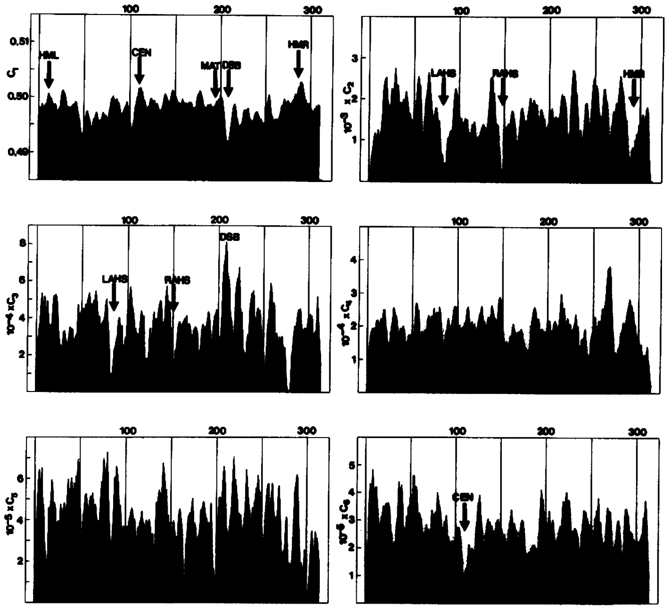

Figure 4. Nucleotide complexity plotted for increasing *n*-tuplets. Details of calculations are given in methods. The values were plotted using a running average of 2000bp. **a.** nucleotides, **b.** dinucleotides, **c.** trinucleotides, **d.** tetranucleotides, **e.** pentanucleotides, **f.** hexanucleotides.

Transcription. The coding regions are characterised as regions of relatively uniform helical stability. These are flanked by regions of lower stability. Opening of the duplex is required for attachment of the transcription machinery. One might imagine that once attached, transcription is facilitated by uniform helical stability. The low helical stability at the 3′ end of most coding regions should be able to accommodate the positive supercoils resulting from unwinding of the helix by RNA polymerases. In a number of cases (102670, 163270, 252120 bp) coding sequences are interrupted by regions of higher stability. The curvature at such points is correspondingly low. Regions 5′ of coding regions tend to have a high density of AT tracts [4] which appears to contribute to the lower stability and reduced curvature.

Transposition. Chromosome III has two regions which have been identified as primary transposition hotspots (LAHS at 87000, and RAHS at 148700bp). A further region has recently been described [26] (FRAHS at 169000bp). These are regions of low stability and locally straight DNA (Figure 1). In addition, both RAHS and LAHS are regions of a U-bend conformation covering several

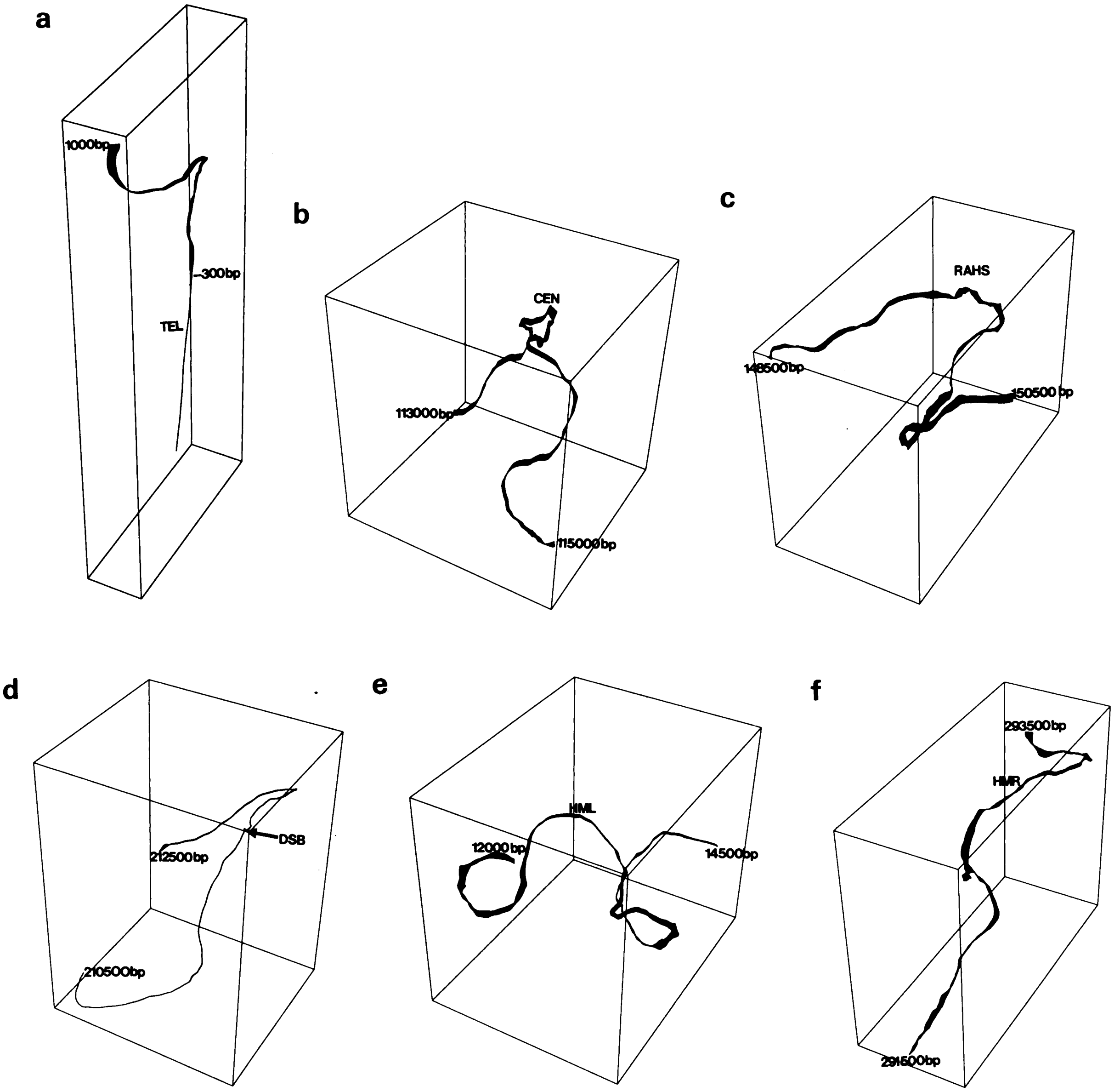

Figure 5. Superimposed three-dimensional path and thermal stability plots for a range of chromosomal features. **a.** left telomere region (TEL), **b.** centromere (CEN), **c.** right arm transpositional hotspot (RAHS), **d.** double-strand break and recombination hotspot (DSB), **e.** *HML* locus, **f.** *HMR* locus.

thousand base pairs, where the hotspots are located in the sides of the U-bend (Figure 5c). As can be seen from Figure 3, the hotspots are very close to each other in three-dimensional space. Spatially, LAHS and RAHS are equivalent to 1630 bp apart, compared to a contour distance of 61700bp, whilst FRAHS is 3000bp equidistant from both LAHS and RAHS, compared to contour distances of 82000 and 20000 bp respectively. The chromosome sequence contains 13 δ elements (of about 333bp), each with single complete Ty2 insertion elements [4]. These appear to be clustered in three regions. The clusters are in non-coding regions of low helical stability (Figures 1; 2), whilst the elements themselves are of relatively uniform stability.

Mating type switching. Three genetic loci are involved in yeast mating type control: *MAT, HML* and *HMR*. Their relative position determined by the global conformation of chromosome III clearly affects their function. *MAT* in the middle of the right arm (198500–200000bp), is the expression locus that determines mating type, whereas *HML* (12500–14000bp) and *HMR* (292000–293000bp) represent silent repositories of mating-type information. The persistent mating type gene rearrangements of this chromosome have previously suggested a higher order structure which preferentially, as required, brings *HML* or *HMR* to the *MAT* locus [1]. Due to shared nucleotide sequence homology at these loci, intrachromosomal recombination can then take place. It can be seen from Figure 3 that the *HMR* and *HML* are located in symmetrically similar positions at the end of subtelomeric regions of relatively low curvature.

The helical contour length from *HML* to *MAT* of 186000 bp is almost exactly twice the 93250 bp from *HMR* to *MAT*. However, taking into account the higher packing ratio between *HML* and *MAT* the spatial distances are almost equal (*HML* to *MAT* of 30166 bp and *HMR* to *MAT* of 32834 bp). This provides structural evidence for the action of the mating type system.

DISCUSSION

The analysis of DNA structural features and sequence complexity along a eukaryotic chromosome has allowed an insight into constraints imposed on sequence construction. The distribution of functional sites is clearly related to structural properties of the DNA at several levels of organisation.

The isochore model [27,28] takes G+C content as being the prime determinant of regional bias. Compartmentalisation of G+C content in this yeast chromosome has already been demonstrated [3]. The data presented here show that whilst this is clearly related to thermal stability of the helix (Figure 1a), over shorter regions the relationship breaks down, due to the cooperative nature of thermal unwinding and the homostabilising propensity of gene-coding regions [7]. Compartmentalisation is thus observed at the level of coding and non-coding regions, with a degree of self-similarity apparent in higher order patterns of stability. As a result, the major constraint on compositional compartmentalisation appears not to be base composition *per se*, rather the propensity of gene coding and some chromosomal features to be of uniform stability. This propensity may contribute to efficient unwinding of the helix for transcription. Insertion of δ sequences by transposition also appears to have a local homostabilising influence.

A number of processes are related to and may cause G+C, and thus stability, variation. The frequency of recombination events or replication timing may have led to the non-uniform distribution in this chromosome [3]. A non-random distribution of possible recombination sites has recently been demonstrated in chromosome III [29]. The experimental strategy involved observing the effects of chromosome fragment plasmids, derived from sites along chromosome III, on meiotic disjunction. The degree of disjunction observed is likely to reflect the propensity of a chromosome site to be involved in pairing of homologous chromosomes. At least two sites had profound effects on self-chromosome specific nondisjunction. One is the 5kbp region proximal to the gene *THR4* in the middle of the right arm, which has also been identified as a hot spot for meiotic recombination (recombining more frequently than most other regions of the genome), as well as containing a preferred site for meiosis-induced double strand breaks in DNA [26]. The recombigenic activity of the region and its influence on chromosome segregation have been shown to be interrelated in this case. The DNA in this region is of elevated stability and has a kink in the DNA path. The effect on the 12 kb region between *HML* and the left telomere is consistent with some earlier cytological studies which showed initiation of pairing and of the synaptonemal complexes at subtelomeric regions within chromosomes in plants, fungi and insects [30,31]. The DNA in this region is also of elevated stability and low curvature. It has been suggested that further analysis with chromosomal fragments would reveal several additional regions along the chromosome [26,29] that are involved in pairing, and that these would be associated with hot spots for meiotic recombination. Candidates for such regions may now be selected, based on DNA structural characteristics.

Replication involves local unwinding of the DNA helix, after recognition of *cis*-acting elements by specific proteins. ARS consensus sequences and active origins have been well studied in yeast chromosome III. It has been demonstrated that there are at least four times as many ARS consensus sequences as would be required to replicate chromosome III. Recent work by Newlon and co-workers [24] has identified a number highly active functional origins, as well as demonstrating that cryptic elements are not activated upon deletion of highly actitve ones. In general, the consensus sequences are shown to be distinguishable by short regions of locally decreased stability, combined with non-curved tracts of the chromosome. In the case of sequences located at 42010 and 78831 bp, which correspond to active elements [4], they are located adjacent to a region of high helical stability (Figure 2). Conversely, a number of inactive elements in the 30kbp distal to the left telomere [4] are located in regions with a lower contrast in helical stability. The level of helicity has previously been found to be quantitatively related to the replication efficiency of cloned ARS mutants [32]. Helical stability also correctly predicted the location and hierarchy of the nuclease-sensitive sites in a C2G1 ARS plasmid [32]. DNA unwinding elements (DUEs) have been established as likely to be necessary to facilitate opening of the origin [10]. The finding that a number of ARS consensus sequences are in regions of a particular three-dimensional conformation adds weight to the argument that replication may be directed by *cis*-acting elements whose accessibility can be limited by the surrounding chromosomal context [33]. Fine detail mapping of the remainding active ARS elements on this chromosome will enable the structural requirements for replication activity to be investigated more precisely.

Yeast chromosomes show interspersion of early- and late-replicating domains, with centromere regions replicating during the first half of S phase and telomeres at the end of S phase [34]. For chromosome III this suggests that regions of lower stability may replicate earlier than more stable regions. In chromosome V there is evidence that the chromosomal context of the origin can affect the time at which the origin is activated [35]. There is also evidence that the unexpressed *HML* and *HMR* loci replicate in the last half of S phase, and that activation of promoters only occurs at the *MAT* locus in an early replicating, and less stable, region of the chromosome. In addition positional silencing of *HMR* requires an ARS [36] (Figure 2).

The three dimensional conformation of the whole chromosome is important for the operation of the mating-type system in yeast. The mating type is determined by donation of a silent cassette from either *HML* or *HMR* to *MAT* by transposition and replication, without loss of the original. Many of the fundamental steps involved in this process may be identical to those involved in the crossing events that exchange blocks of genes between homologous chromosomes during recombination. The large discrepancy in DNA contour length between the two silent loci and *MAT* is now resolved by the finding that the spatial distances are equivalent. The unusual conformation at *MAT* and *HML* loci provides some insight as to how the initial recognition and pairing might take place.

It has been suggested [4] that the 122kb (168261–290110) which is free of δ elements may be resistant to Ty transpositional insertions, or that occurrences of δ or Ty are more deleterious than in other regions. This region has a lower packing ratio than other regions, with correspondingly more stable sequences. However, the finding that the transposition hotspots are proximal in three-dimensional space despite being located on different chromosome arms suggests that transposition is more likely to occur in a particular chromosomal region and conformation.

The three-dimensional path of the DNA helix is thus clearly important in determining the availability of sequences for processes such as recombination and transposition. Whilst there is no overall relationship between thermal stability and DNA curvature, regions of particularly straight DNA tend to be correspondingly more stable. The large-scale patterns of curvature and stability appear to influence the sites of recombination.

The three-dimensional structure and stability is determined by the cooperative effects of particular subsets of nucleotides. As such, the choice of sequences used in coding and non-coding regions is non-random and constrained. As might be expected in a chromosome highly enriched with gene-coding sequences, the information content as measured by 'complexity' is relatively high throughout. There is evidence of overlapping compartments delineated by order at base level up to hexanucleotide level. These correspond to many of the chromosomal features described above. A previous study [37] of *n*-tuple composition, using Markov chain analysis of 392kbp of yeast DNA sequences, demonstrated that tetranucleotide frequencies could be used in predicting hexanucleotide frequencies, and that there was a strong dependence of oligomer frequencies on base composition. This implies that most complexity or information content is accounted for by *n*-tuples of one to four, as is shown in Figure 4. The distributions of sequence complexities appear not to be related to the distribution of gene-coding/non-coding sequences, reflecting the observed secondary effect of underlying chromosomal context on codon usage [3].

The overall picture of the chromosome obtained by plotting the DNA path based on the wedge model is striking in its degree of compaction, symmetry and identification of telomere and subtelomeric regions. Self similarity is observed in the three dimensional path, as well as DNA stability. It is interesting to note here that the DNA path plotted as a 2-dimensional projection displays remarkable qualitative similarity to an electron micrograph of a 300kbp replicating yeast DNA molecule in Fig. 1 of reference [32].

The data from this complete chromosome suggest that the mosaic structure of eukaryote genomes is preserved at several levels of organisation, lending weight to Takahashi's [38] fractal model of chromosomes and chromosome replication, which predicts several levels of self-similarity in coiling and other structure.

ACKNOWLEDGEMENTS

I am particularly grateful to Dr James Lynn for his help in plotting the three-dimensional DNA paths.

This work was funded by the OST through the Agricultural and Food Research Council.

REFERENCES

1. Oliver, S.G. et al. (1992) *Nature* **357**, 38–46.
2. Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R. and Sonnhammer, E. (1992) *Protein Science* **1**, 1677–1690.
3. Sharp, P.M. and Lloyd, A.T. (1993) *Nucleic Acids Res.* **21**, 179–183.
4. Karlin, S., Blaisdell, B.E., Sapolsky, R.J., Cardon, L. and Burge, C. (1993) *Nucleic Acids Res.* **21**, 703–711.
5. Gotoh, O. (1983) *Adv. Biophys.* **16**, 1–52.
6. Wada, A. and Suyama, A. (1985) *Prog. Clin. Biol. Res.* **172**, 37–46.
7. Wada, A. and Suyama, A. (1986) *Prog. Biophys. Mol. Biol.* **47**, 113–157.
8. Wada, A. and Suyama, A. (1986) In Bishop, A.R. and Kawabata, C. (eds), Proceedings of the International symposium on Computer Analysis for Life Science, Ohmsha Ltd, pp140–150.
9. Breslaur, K.J., Frank, R., Blocker, H. and Marky, L.A. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 3746–3750.
10. Natale, D.A., Umek, R.M. and Kowalski, D. (1993) *Nucleic Acids Res* **21**, 555–560.
11. Ulanovsky, L.E. and Trifonov, E.N. (1987) *Nature* **326**, 720–722.
12. Turnell, W.G. and Travers, A.A. (1992) *Methods Enzymol.* **212**, 387–399.
13. Shannon, C.E. and Weaver, W. (1949) The mathematical theory of communication. University of Illinois Press, Urbana.
14. Gatlin, L.L. (1972) Information theory and the living system. Columbia Press, New York.
15. Shenkin, P., Erman, B. and Mastrandrea, L.D. (1991) *Proteins Struct. Funct. Genet.* **11**, 297–313.
16. Lauc, G., Ilic, I. and Heffer-Lauc, M. (1992) *Biophys. Chem.* **42**, 7–11.
17. Sibbald, P.R., Banerjee, S. and Maze, J. (1989) *J. Theor. Biol.* **136**, 475–483.
18. Lerman,L.S. and Silverstein, K. (1987) *Methods Enzymol.* **155**, 482–501.
19. Fixman, M. and Friere, J.J. (1977) *Biopolymers* **16**, 2693–2704.
20. Gotoh, O. and Tagashira, Y. (1981) *Biopolymers* **20**, 1033–1042.
21. Benight, A.S. and Wartell, R.M. (1983) *Biopolymers* **22**, 1409–1425.
22. Trifonov, E.N. and Ulanovsky, L.E. (1987) In Wells, R.D. and Harvey, S.C. (eds), Unusual DNA Structures. Springer, NewYork, pp. 173–187.
23. Bolshoy,A., McNamara, P., Harrington, R.E. and Trifoniv, E.N. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 2312–2316.
24. Dershowitz, A. and Newlon, C.S. (1993) *Mol. Cell. Biol.* **13**, 391–398.
25. Goldway, M., Sherman, A., Zenvirth, D., Arbel, T. and Simchen, G. (1993) *Genetics* **133**, 159–169.
26. Oliver, S.G., James, C.M., Gent, M.E. & Inge, K.J. (1993) In Heslop-Harrison, J.S. and Flavell, R.B. (eds), The Chromosome. BIOS Scientific Publishers, Oxford, pp. 233–248.
27. Bernardi, G. and Bernardi, G. (1986) *J. Mol. Evol.* **24**, 1–11.
28. Matassi, G., Montero, L.M., Salinas, J. and Bernardi, G. (1989) *Nucleic Acids Res.* **17**, 5273–5290.
29. Goldway, M., Arbel, T. and Simchen, G. (1993) *Genetics* **133**, 149–158.
30. Holm, P.B. and Rasmussen, S.W. (1980) *Carlsberg Res. Commun.* **46**, 300–346.
31. Albini, S.M. and Jones, G.H. (1987) *Chromosoma* **95**, 324–338.
32. Natale, D.A., Schubert, A.E. and Kowalski, D. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 2654–2658.
33. Fangman, W.L. and Brewer, B.J. (1992) *Cell* **71**, 363–366.
34. McCarroll, R.M. and Fangman, W.L. (1988) *Cell* **54**, 505–513.
35. Ferguson, B.M. and Fangman, W.L. (1992) *Cell* **68**, 333–339.
36. Rivier, D.H. and Rine, J. (1992) *Trends Genet.* **8**, 169–174.
37. Arnold, J., Cuticchia, A.J., Newsome, D.A., Jennings, W.W. and Ivarie, R. (1988) *Nucleic Acids Res.* **16**, 7145–7158.
38. Takahashi, M. (1989) *J.Theor. Biol.* **141**, 117–136.