Progr Colloid Polym Sci (2006) 131: 97–107
DOI 10.1007/2882_009

Published online: 15 February 2006

BIOLOGICAL SYSTEMS

Oliver Clay
Nicolas Carels
Christophe J. Douady
Giorgio Bernardi

Density Gradient Ultracentrifugation and Whole Genome Sequences: Fine-tuning the Correspondence

Oliver Clay · Nicolas Carels · Giorgio Bernardi (✉)
Laboratory of Molecular Evolution, Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Napoli, Italy
e-mail: bernardi@szn.it

Christophe J. Douady
Équipe d'Hydrobiologie et Ecologie Souterraines & Plateforme d'Ecologie Moléculaire, Laboratoire d'Ecologie des Hydrosystèmes Fluviaux, Université Claude Bernard Lyon 1, UMR CNRS 5023, 69622 Villeurbanne Cedex, France

Abstract Since its introduction in the 1950's, analytical ultracentrifugation (AUC) of DNA in CsCl and other salt density gradients at sedimentation equilibrium has remained an elegant way to gain insight into the variation of base composition (GC, guanine + cytosine %) among and within animal and plant chromosomes, and into functional correlates of GC. Absorbance profiles of routine preparations of DNA in CsCl are essentially GC histograms of fixed-length sequence fragments (≈ 15–100 kb). This correspondence has been amply illustrated by genome sequences obtained over the past 5 years. Both AUC and sequencing have now generated large amounts of data that can be jointly mined. The dialogue between these two approaches should render tractable some tenacious problems of CsCl profile analysis, such as the correct treatment of concentration dependence for heterogeneous DNA. We focus on how absorbance profiles of a species' DNA vary as one changes the scale of one's observation (molecular weight), and dissect this scale-dependence into the contributions from its two main sources (diffusion, sequence effects). Our understanding of heterogeneous DNA in CsCl gradients can profit from the comparison of results from AUC and whole-genome sequencing, and the insights gained should prompt more strategic AUC analyses of DNA.

Keywords Analytical ultracentrifugation · Sedimentation equilibrium · Base composition · Evolution · Long-range correlations

Introduction

The base composition of a DNA molecule is primarily its GC level, the molar ratio of GC (guanine-cytosine) base pairs in the DNA. If one considers just one strand, GC is the percentage of nucleotides that are G or C, and not A or T. This is a most fundamental property of DNA. Numerous functional and evolutionary correlates of its variation along chromosomes, in taxa ranging from bacteria to human, are now known [1].

CsCl gradient density ultracentrifugation of DNA was introduced in 1957. Its principle is well summarized in the original paper [2]: "*A solution of a low-molecular weight solute [e.g., CsCl] is centrifuged until equilibrium is closely approached, [resulting] in a continuously increasing density along the direction of centrifugal force. Consider the distribution of a small amount of a single macromolecular species [e.g., DNA] in this density gradient. The initial concentration of the low-molecular-weight solute, the centrifugal field strength, and the length of the liquid column may be chosen so that the range of density at equilibrium encompasses the [buoyant] density of the macromolecular material. The centrifugal field tends to drive the macromolecules into the region where the sum of the forces acting on a given molecule is zero. (The [buoyant] density of the macromolecular material is here*

defined as the density of the solution in this region.) This concentrating tendency is opposed by Brownian motion, with the result that at equilibrium the macromolecules are distributed with respect to concentration in a band of width inversely related to their molecular weight."

Soon after the introduction of this technique, which was conceived with labelling in mind (see [3] for a historical account), it was discovered that there is an important, simple and accurate empirical link between analytical ultracentrifugation (AUC) and GC [4–6]. This link has been used routinely since then, yet it still awaits a full quantitative, physicochemical explanation. In a CsCl density gradient, at sedimentation equilibrium, the GC of a macromolecule of DNA is linearly related to its (time-averaged) buoyant density. The buoyant density of that DNA is, in turn, practically a linear function of its radial position: the density gradient is effectively linear over the distances from the ultracentrifuge axis that are of interest. The positional distribution of the DNA macromolecules in a gradient is scanned by an analytical ultracentrifuge and reported as an absorbance profile. A DNA sample consisting of molecules or fragments that have different base compositions (GC) will have, at sedimentation equilibrium, a profile of finite width. To a good approximation, valid for most species' DNA and especially if the molecular weight is high, this equilibrium profile is just the GC distribution of the molecules.

Corrections that improve the accuracy of the match are based on tractable, well-known effects experienced by macromolecules in solution. Indeed, there are basically three components: DNA, salt, and water, although DNA macromolecules can have different sequences that make them behave in solution, or interact, in different ways by adopting anomalous configurations or aggregating. The problem is not so much in understanding the individual effects involved, but in assessing their relative importance and cross-influencing, which encumbers a modular treatment. Subtle but persistent impediments to perfect matching may involve DNA concentration effects and/or aggregation, anomalous sedimentation of repetitive DNA, and methylation.

Molecular weight polydispersity, where present, can add further complexity, although in many situations it is unproblematic: current DNA extraction protocols typically produce narrow molecular weight distributions, so in calculations one can simply use the mean.

GC and its contrasts are of interest because they correspond to functionally and evolutionarily telling genome properties. For example, in mammals and birds the GC-richest regions of a genome have the highest gene densities and expression levels, the most interior locations in the nucleus at interphase and the earliest replication in S-phase, preferentially open chromatin and short-intron genes, and more frequent and longer CpG islands (reviewed in [1]).

The analytical ultracentrifuge is likely to soon become more refined, precise and versatile (see, e.g., [7]), and whole-genome sequencing should become cheaper and faster (see [8–10]). Such technical advances promise to allow more accurate and varied comparisons between GC distributions of genomic DNA sequences and their CsCl absorbance profiles. It now seems, therefore, the right time to address remaining open problems in our understanding of how heterogeneous DNA behaves while it is being ultracentrifuged in CsCl gradients. In this paper, we use results from density gradient AUC and entirely sequenced genomes to specify how sequences and absorbance profiles should dovetail. Proper dovetailing permits consistency checks that can tell us if our ideas about profile formation are correct.

Experimental

At sedimentation equilibrium in a CsCl density gradient, the GC level of a molecule or fragment of double-stranded DNA is related, with a few exceptions that are listed below, to its (time-averaged) buoyant density ρ in the gradient [4–6]. The linear equation that relates these two quantities is [11]

$$\mathrm{GC} = \frac{\rho - 1.660\,\mathrm{g\,cm^{-3}}}{0.098} \times 100\%\,. \tag{1}$$

Because there are exceptions to this rule, buoyant density (also called "density" in early publications), rather than GC, is often chosen as the quantity of interest in AUC studies. GC is, however, the ultimate object of our investigations.

Buoyant density in CsCl is, in turn, a simple function of the distance r of the molecule from the axis of an analytical ultracentrifuge [12],

$$\rho = \rho_\mathrm{m} + \kappa\omega^2(r^2 - r_\mathrm{m}^2)\,. \tag{2}$$

Here, ρ_m and r_m are, respectively, the buoyant density in CsCl and radial position of a suitable marker, such as bacteriophage 2C (which has a very high ρ_m, $1.742\,\mathrm{g/cm^3}$ because of its modified bases). ω is the angular speed and κ is a constant that depends on the details of the ultracentrifuge cell and the rotor ($\kappa \approx 4.2 \times 10^{-10}$ for Beckman models E and XL-A). Since the differences in radial position are very small compared to the distances from the axis, we can write $r^2 - r_\mathrm{m}^2 \approx 2r_\mathrm{m}(r - r_\mathrm{m})$, i.e., the nonlinearity in Eq. 2 is negligible, and the gradient is essentially linear.

For a generic, high molecular weight DNA sample (e.g., 50–100 kilobases or kb) in which we can neglect diffusion of the molecules or fragments, their radial distribution in the CsCl gradient is, after a linear calibration, just their GC distribution. At lower molecular weights, diffusion visibly broadens the radial distribution. Indeed, small DNA molecules or fragments will continue to diffuse appreciably around their expected positions in the gradient at sedimentation equilibrium, although the average number of molecules or fragments at any given position will not

change, i.e., at a macroscopic scale there will be no more net motion.

The absorbance profile that reports the radial distribution thus corresponds, via Eqs. 1 and 2, to the GC distribution of the DNA molecules, if their molecular weight is high and they do not contain satellite or modified DNA. For this correspondence to hold, however, certain standard experimental conditions must be met (see, e.g., [13] and [14] for partial lists). Among the conditions, we mention that the average concentration should be neither too low, because the absorbance of DNA is then easily confounded with a possibly irregular "baseline" that may include minor contaminants, nor too high, because the response is no longer linear when DNA crowds excessively at the band center. If CsCl solution is excluded from the region and/or light is blocked, this can locally deform the otherwise linear gradient and/or flatten the profile. Absorbances are measured at 260 nm and the standard speed is 44 000 rpm.

The Shaping of the DNA Absorbance Profile: Physicochemical Contributions

We begin with a historical absorbance profile: the first view of a mammalian genome's GC distribution ever published. Figure 1 shows the positional distribution of fragments of DNA from a calf's thymus DNA in a CsCl gradient, at sedimentation equilibrium, reproduced from the original paper that introduced density gradient ultracentrifugation [2]. The speed of the rotor (44 700 rpm) was similar to those used in routine CsCl analyses today; only the molecular weight (presumably around 5 kb) was not as high. Such equilibrium profiles can be obtained after about 24 hours. An introduction to the method's physical foundations can be found in the book by van Holde et al. [15].

The caption accompanying the 1957 plot concluded with the sentence: "The skewness in the resultant band indicates heterogeneity in effective density". Such marked heterogeneity of the effective density, or buoyant density, is generally pronounced in mammals and birds, and results largely from GC heterogeneity among the chromosomal fragments represented in the sample. In the special case of cow, the profile heterogeneity is further exaggerated by the very high percentage ($\approx 25\%$) of highly repetitive satellite DNA in the bovine genome, as was discovered later [16].

In October 2004, the first draft of the cow genome sequence was placed in the public domain (http://genome.gov/12512874). The full assembly of the large scaffolds and the publication describing the sequence have not yet appeared. If one allows for some inaccuracies due to gaps, the GC levels of the sequenced Hereford cow's 5 kb segments (or, rather, of the presently available scaffolds' segments) can be fetched from an annotation database that now exists for this genome and plotted as a histogram, as is shown in Fig. 2. In this way we can again see the cow genome's GC distribution, 47 years and some $53 million after the scan of Fig. 1.

We now go into technical detail. When one overlays (after converting to GC units) an experimental curve such as the one in Fig. 1, which represents collections of similarly sized fragments of a genome, by its sequence-derived counterpart, such as the histogram in Fig. 2, one finds that the absorbance profile is wider than the histogram. The two main reasons for this difference in cow are well known: highly repetitive DNA [16], and diffusion.

Highly repetitive DNA is present only in modest amounts in sequence scaffolds: heterochromatic regions such as centromeres are not (and will not soon be) tar-

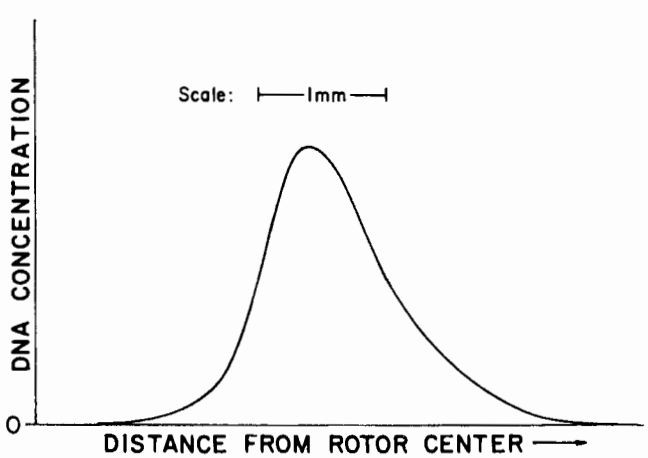

Fig. 1 Equilibrium profile of the cow genome in CsCl, obtained in 1957 when CsCl gradient AUC was introduced. Molecular weight is presumably less than or around 5 kb, other conditions are standard. Reproduced with permission from Meselson et al. [2]

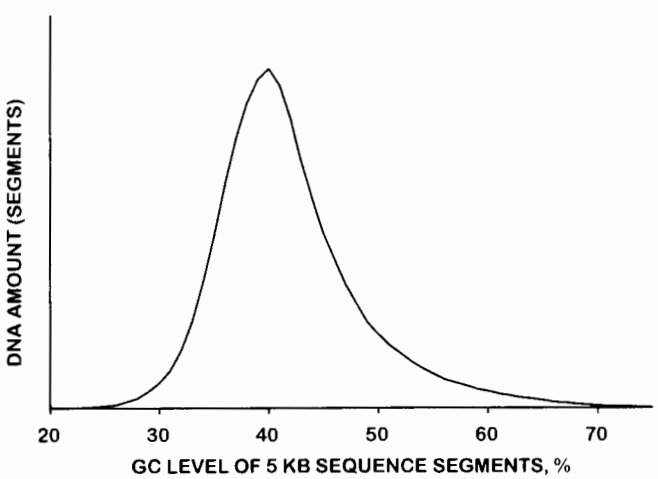

Fig. 2 GC histogram of 5 kb segments of the recently sequenced cow genome. The draft sequence's GC levels were obtained from the gc5base.txt table of the UCSC annotation database for cow, http://genome.ucsc.edu (ratio sumData/validCount). Bins of 1% GC were used

geted for sequencing, since they rarely contain genes and are more difficult to sequence. Where scaffolds do contain such DNA, the sequences report the correct GC. In an ultracentrifuged sample, on the other hand, a fair number of the molecules will consist of (largely unsequenced) highly repetitive DNA, which can extend over long tracts of unstable length, expanding or contracting even from one generation to the next as a result of mechanisms such as slippage and out-of-register homologous recombination. Such satellite DNA will often (but not always) appear as visible peaks or bumps within or next to the main-band profile: since the many molecules of a particular satellite are identical or almost identical, their GC level will be overrepresented in the absorbance profile. The amount of a particular satellite in a genome, i.e., the height of its peak or bump in the profile, will often change quickly over evolutionary timescales, sometimes differing visibly within genera (as in kangaroo rats [17]). Some of the repetitive satellite DNA may also band anomalously, i.e., be found at another position in the gradient than expected on the basis of its GC (as in guinea pig [18]). Such abnormal banding occurs presumably because the repetitiveness (and/or methylation) of the satellite fragments alters their buoyant density.

Diffusion is an effect that is irrelevant to GC distributions obtained from sequences, whereas in a CsCl gradient the DNA molecules undergo random Brownian motion around their expected equilibrium positions, and thus broaden the absorbance profile. When molecules or fragments are identical, diffusion broadens the profile in inverse square proportion to the fragments' lengths.

To approach an understanding of how diffusion acts on DNA in real situations, we begin by considering the natural DNA that gives the simplest CsCl profile: a preparation of intact, identical copies of a bacteriophage such as phage lambda. Such a sample of lambda DNA will have no intermolecular GC heterogeneity (the molecules will typically all have the same sequence, gb:lamcg), and no polydispersity in molecular weight ($M = 48.5$ kb). Because of the simplicity of the scenario they permit, phages have been often used as models for studying macromolecules' behavior in density gradients. The absorbance profile of such a homogeneous, monodisperse DNA sample will report a Gaussian distribution of molecules' positions. The width (standard deviation) of that Gaussian profile will depend on diffusion, but also to some extent on the amount of DNA that was loaded into the ultracentrifuge cell: virial effects will cause an additional broadening of the profile when the overall DNA concentration is high. The virial effect is absent, by definition, at infinite dilution, and in phages it has been observed to increase exponentially with increasing concentration. We can therefore summarize the situation for homogeneous DNA samples by saying that their absorbance profiles are well described by a Gaussian distribution with standard deviation $\sigma = \sqrt{a/M}\,e^{Bc}$, where M is the molecular weight, c is a measure of overall concentration or of the amount of DNA loaded, such as the maximum absorbance (optical density), a is a factor that takes solvation into account, and B can depend on the species analyzed [19, 20].

Vertebrate genomes or chromosomes are much larger than those of intact phages, and are therefore represented only by their fragments in DNA samples: during the experimental preparation of DNA, a very large macromolecule such as a cow chromosome is inevitably sheared into fragments. Where a genome is heterogeneous in GC, there will then be intermolecular GC heterogeneity in the sample, and thus broad profiles, which will be further broadened by diffusion and/or satellites. Local narrowing can also occur: if some but not all DNA aggregates during the approach to equilibrium, the aggregating DNA will attain a higher molecular weight than the rest of the DNA, so an effective polydispersity can develop, with the aggregating DNA forming a more highly peaked "subprofile" (see [21] for an example in mouse). Several other factors can also play a role in shaping or shifting profiles, such as DNA methylation, electric effects, pressure effects, local deformations of the CsCl gradient where DNA is crowded, or light bending by the gradient. Most of these factors hardly distort the relatively broad profiles of vertebrates.

At infinite dilution, there is no concentration dependence, and the total profile variance of a heterogeneous sample is then the GC distribution's variance plus the diffusion variance $\sigma^2_{\text{diffusion}} = a/M$ [20]. In other words, we have a convolution of an (often non-Gaussian) distribution representing the GC heterogeneity and a Gaussian distribution of unit area (also called a filter, kernel, or point spread function) representing the diffusion broadening. Once the constant a is known, the standard deviation of the GC distribution can be calculated from the absorbance profile and an estimate of the sample's molecular weight M. With some inevitable numerical inaccuracies, we can then even extract the full GC distribution by deconvolving, i.e., "peeling off" the Gaussian of unit area that represents the diffusion. In real situations, dilutions are finite, and when we wish to take concentration effects into account we can no longer carry over the solution from the homogeneous case: virial effects and possible aggregation can act simultaneously but in opposite directions (broadening or narrowing the profile) and in different parts of the profile (tails or center).

The remarks we have made so far pertain to ways in which DNA macromolecules' behavior in a density gradient shape their absorbance profile. The shaping entails a molecular weight dependence that is especially strong for small fragments, but becomes weak or negligible for long fragments. It has been understood in its rudiments for over three decades. This rudimentary understanding often permits accurate extraction of the underlying GC distribution. In most cases where the entire genome sequence is known, one finds that by neglecting concentration de-

pendence, and by assuming a Gaussian point spread function whose width is inversely related to the square root of the fragments' mean length, one obtains estimates of the underlying GC distribution that differ only very slightly from the sequence-derived GC distribution (cf. [14]). In some species or conditions, however, the profile and/or its concentration dependence depart visibly from expectations.

Another contribution to the profile's final shape is of practical as well as molecular-biological and evolutionary interest: the GC distribution, which entails a molecular weight dependence that is intrinsic to the genome under study.

The Shaping of the DNA Absorbance Profile: GC Contributions

At the sequence level, molecular weight corresponds to fragment length or, if one views the fragment as part of a chromosome sequence, to segment or window length. Thus a GC distribution that has been deduced from an absorbance profile will correspond to a sequence-derived GC histogram. If the sample consists of heterogeneous but monodisperse fragments of chromosomal DNA, the histogram will be a histogram of fixed-length segments or windows along the chromosome(s).

Sequence heterogeneity can be investigated experimentally by CsCl gradient ultracentrifugation over two orders of magnitude ($\approx$ 3–300 kb), using a simple principle: intramolecular heterogeneity becomes intermolecular heterogeneity when one regards smaller molecules or fragments.

Thus, much of the intragenomic GC heterogeneity in mammals was well understood already in the 1970's, even though no chromosomal regions (and only a handful of genes) had known sequences at that time. The investigations using AUC led to the following picture. Sequences of mammalian DNA (GC and AT) do not resemble runs of independent coin tosses (heads and tails), but have a remarkable organization. Within relatively short (kb) regions, GC levels are already long-range correlated. At larger scales (> 100 kb), GC levels are organized in a mosaic as one travels along a chromosome, in which GC-rich segments alternate with GC-poorer segments. The pieces of the mosaic, which are much more homogeneous in GC than the whole genome and persist in this relative homogeneity over long distances, from $\approx$ 300 kb to several megabases, are called isochores ([21–23]; their discovery, functional correlates and evolution are reviewed in [1]).

Intragenomic heterogeneity at a given scale, such as 10 or 70 kb, can be quantified by the width of the CsCl absorbance profile or, more precisely, by the standard deviation of its underlying GC distribution. Thus, plotting the width (standard deviation) of a genome's GC distributions versus the relevant segment size (molecular weight) gives an immediate visual impression of that genome's GC heterogeneity at all scales. Plots of this type are shown in Fig. 3, on double-logarithmic scales. The slope of the log-log plot for human is around -0.15 (already much less steep than for an uncorrelated sequence) at low molecular weights < 10 kb, and then gradually flattens out, finally reaching a horizontal plateau that remains constant for molecular weights greater than about 70–100 kb. In other words, the narrowing of a mammalian CsCl profile, as molecular weights are increased, slows down and reaches a constant width from about 70–100 kb onwards. In fact, at 100 kb a mammalian profile (and not just its standard deviation) is practically indistinguishable from one at 300 kb. This observation was first made using AUC [21] and was recently confirmed with genome sequences [24]. The original observation was a key element in deducing the existence of isochores [21], since it is exactly what one would expect if chromosomes are organized into long regions $\gg$ 300 kb within which heterogeneity is remarkably low compared to the genome-wide heterogeneity, and it is very far from what one would expect if no similar large-scale structure were present.

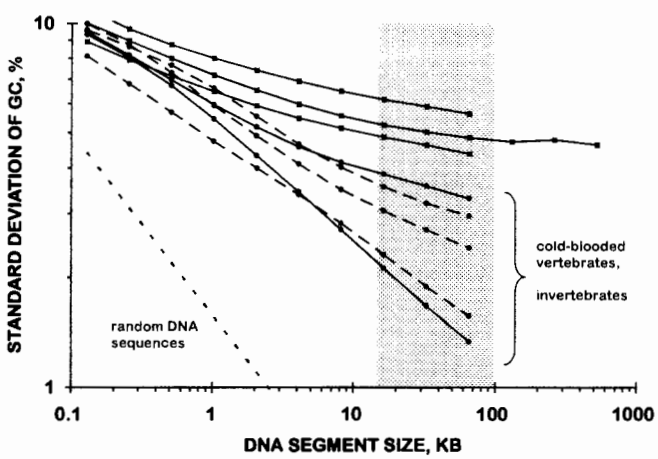

Fig. 3 GC standard deviations of entirely sequenced animals at different molecular weights. Vertebrates are shown by *solid lines*, from *top to bottom*: dog, human, chicken (*square markers*), *Tetraodon* pufferfish, zebrafish (*circular markers*). Protostomes (*dashed lines* and *circular markers*, from *top to bottom*): *Anopheles gambiae* mosquito, *Drosophila yakuba* fly, and *Caenorhabditis elegans* nematode worm. Sequences were retrieved from the UCSC genome browser (http://genome.ucsc.edu, GoldFasta files). Logarithmic scales are used for both axes, since both random DNA (*dotted line at bottom left*, log interval slope of $-1/2$) and power-law correlated DNA (slope less steep than $-1/2$) give straight lines on such plots. "Complete" vertebrate sequences still contain gaps, including ribosomal and heterochromatic DNA, so the plotted values are approximate. The *grey region* indicates the interval (15–100 kb) in which most AUC data points will be found. For clarity only the human plot is extended past this interval

Implications of Long-range Correlations in Isochores

The detection and characterization of long-range correlations or "long memory" in sequences or time series is an active field of research that has interested not just biologists, hydrologists, physicists, mathematicians and statisticians, but also econometrists and financial modellers, and several methods exist for estimating such correlations' characteristic parameters (see, e.g., [25] or Chapter 8 of [26]). We here mention only the detection and estimation method that is natural in the context of CsCl gradient AUC. Long-range positive, serial autocorrelations in GC that are present in DNA can be detected by a linear decrease in the log-log plot of inter-segment standard deviation versus segment length, with a slope that is distinctly less steep than $-1/2$ over one or two orders of magnitude (see, e.g., [25, 27–29]). A slope of $-1/2$ is what one would expect for a random sequence of uncorrelated or independent nucleotides; a less steep slope but again following a straight line is what one would expect for a serial autocorrelation that decreases as a power of the intervening distance d (proportional to $d^{-\alpha}$). The closer the slope is to 0, the higher is the long-memory parameter, i.e., the more serious is the departure from a statistical scenario of independent and identically distributed (i.i.d.), or random nucleotides. In particular, no short-range (Markovian) dependence can be assumed in such cases: the dependence is irreparably long-range.

The fact that one cannot invoke familiar textbook scenarios has far-reaching implications. For one thing, the independence assumption is a pillar on which most of traditional statistics is based. When this assumption falls, many standard tests for large-scale DNA properties or contrasts become invalid, and the modified tests that are applicable (see Section 8.6 of [25] for examples) have lower apparent statistical power, at a given length scale. Similarly, traditional models used to reconstruct phylogenies from sequences typically assume independence among nucleotides.

Using only analytical ultracentrifugation in CsCl, one can already use standard deviations of a species' profiles, obtained at different molecular weights, to obtain log-log plots such as those shown in Fig. 3, and then use the slopes to estimate bulk long-range memory parameters (e.g., an estimate of the Hurst exponent H is the slope plus 1). One can, furthermore, experimentally obtain compositional fractions or Gaussian components [23]. By plotting their standard deviations individually against molecular weight, one observes that the standard deviation and long-memory parameters (non-independence) of mammalian DNA are systematically higher for GC-rich isochores than for GC-poor isochores, at molecular weights up to around 70 kb. All of this can be done in principle, and largely also in practice, without knowing any sequences ([22, 23]; see [27] for a discussion of the concordance with sequence results).

One consequence of long-range dependence that is directly visible during routine sequence analyses is statistical self-similarity, which holds in the "linear" range of the log-log plot, i.e., at scales up to about 70–100 kb. Indeed, moving-window GC scans of chromosomal regions obtained using one window size look, statistically, very similar to those obtained using a much smaller or much larger window size, if one correctly resizes the vertical axis (which one can do by consulting the standard deviation versus window plot). Thus, it is often impossible to deduce anything about the scale (window size/molecular weight) by just looking at the scans.

Comparing Species when the Molecular Weights are Different

The general features described above are common to eutherian mammals, but the quantitative details can be quite different among individual species. If we had a collection of DNA samples from different mammals, all samples having the same molecular weight, we would have main band profiles with different modes, means, standard deviations (as shown in Fig. 3) and/or asymmetries. In other words, different mammals have different GC distributions. These differences can be phylogenetically informative, and/or tell us about base compositional shifts that occurred during mammalian evolution, such as the narrowing of the profile in a rodent lineage that led to mouse and rat.

We would often like to compare different species' genomes by comparing their GC distributions. We may have one DNA sample for each of the species, but then find that those samples' molecular weights are different. This situation is not uncommon, since samples collected in the wild, preserved under different conditions by different investigators, and injected into the ultracentrifuge cell can end up with average molecular weights varying from about 10 kb (or less) to 100 kb. If we can accurately measure the average molecular weight of each sample, for example by pulsed-field gel electrophoresis, we can then estimate the GC distributions and/or their standard deviations.

As mentioned above, a GC distribution of fixed-length fragments or segments of mammalian DNA (< 100 kb) will be narrower if the fragments are long than if they are short. Because of this molecular weight dependence, it can be difficult to compare profiles from two different species if one species is represented by, say, a 70 kb sample and the other species is represented only by a 10 kb sample. Ideally, one would like to predict what a profile would look like if the same species' sample had instead some other, standard molecular weight. The plots of Fig. 3, obtained from recently sequenced genomes, show that this is in general a difficult task. Indeed, for mammals, other vertebrates, or even protostomes such as insects and ne-

matodes there is no general rule that would allow us to reliably "convert" a species' profile obtained at 10 kb to the same species' profile at 70 kb. Conversely we might be tempted to shear the 70 kb down to about 10 kb and then centrifuge the smaller fragments, but Fig. 3 shows that interspecific comparisons at 10 kb would no longer allow much resolution: valuable large-scale information that can differ markedly among genomes is destroyed by such shearing. The best we can offer, to shed some light on the possible molecular weight dependencies of different taxa, is a "calibration" of the plot of standard deviation versus molecular weight: one traces the molecular weight dependencies for species represented by several samples of different molecular weights [21, 23] or by a whole-genome sequence.

Both the utility and the limitation of such a calibration are seen in Fig. 3. In the molecular weight range that is of most interest for CsCl work ($\approx$ 15–100 kb), it is a promising sign that the lines traced by different vertebrate species do not cross, but fan out. In other words, if we have one fish genome's GC distribution for only 10 kb fragments and another fish genome's GC distribution for only 70 kb, both obtained via ultracentrifugation, then we can guess (but only guess, until we have more sequence- or AUC-derived traces of fishes as calibrating guidelines) the angle at which the traces of Fig. 3 would pass through each of those two data points.

The fact that zebrafish has narrower profiles than pufferfish, at a given molecular weight > 1 kb, is of interest since the genomes of these two bony fishes are comparable, and indeed the difference between their heterogeneities has long been known from AUC (see [30] and references therein). Similarly, the observation that the warm-blooded vertebrates have distinctly higher GC heterogeneities than the cold-blooded vertebrates exemplifies a well-documented difference of functional and evolutionary significance, and several lines of reasoning point to an explanation in terms of thermal stability ([31]; reviewed in [1]). The difference between the two dipteran insects, mosquito and fly, is also interesting, since they have diverged considerably in their genomes' compositional properties [32], during the 250 million years or so since their lineages separated [33].

Figure 3 also shows that there is no hierarchy separating all vertebrates from all insects or nematodes: some large-scale genomic GC differences obviously evolved independently in fishes and insects. This example illustrates the danger of excessively widening the taxonomic range of one's comparison: a homoplasy, i.e., a convergent or apparently convergent evolution of a trait (in this case GC heterogeneity) will mislead phylogenetic reconstructions if they are based on that trait. Zebrafish is obviously not more related to a worm than to pufferfish: the evolutionary distances over which one is trying to directly compare GC data are too wide. To reduce the chances of being misled, one can include the profile mean or mode as a second trait (giving a 2D plot, if molecular weights are similar; see [34] for an example), restrict one's phylogenetic range, and sample taxa densely within that range.

Concentration Dependence and the Approach to Equilibrium

The good agreement, after applying established corrections, between CsCl absorbance profiles and genome sequence-based histograms also indicates where theoretical improvements would be welcome. One remaining unsolved problem is how heterogeneous equilibrium profiles change as different amounts, i.e., concentrations of DNA are loaded. Another interesting problem would be to describe how heterogeneous pre-equilibrium profiles change as they approach equilibrium. A solution to the first problem, if indeed a generally applicable solution exists, would be of much help for extracting accurate GC distributions from AUC profiles. A solution to the second problem might be useful in providing a rough but "real-time" estimate of a sample's molecular weight, without needing to resort to separate pulsed-field gels or sedimentation velocity runs, or for double-checking molecular weight estimates obtained via these other routes (see Appendix). It would be particularly useful for detecting and monitoring, in real time, any unexpected aggregation of DNA into high-molecular weight clusters, for example as a satellite band or crowded main band center is being formed.

Not all calculations for homogeneous samples can be easily transferred to heterogeneous samples, and indeed the exponential concentration dependence of homogeneous DNA cannot be simply ported. Aggregation of DNA is often likely to enter the picture: it narrows (instead of broadening) a profile when DNA concentration is raised. In fact, if only the virial effect, observed for phages, were active, one might be tempted to try a folding (generalized convolution) operation, in which again a Gaussian spread function, this time with a width that increases exponentially with (local) concentration, spreads a heterogeneous genome's GC distribution. The most flattened region of the profile would then be the modal region near the band center. This is not observed, and instead the modal region is often narrower than one would expect from the corresponding genome sequence. It is not yet clear if such observations are generally best explained by aggregation of crowded DNA, or if they are almost as often caused by other effects, or by unsequenced satellite DNA that may be present in the profile but not in the sequence-derived GC distribution. For a complete treatment one must probably return to first principles.

Concerning the second problem, the approach to equilibrium of heterogeneous DNA, one might begin by considering a simple "toy" genome model consisting of two

well-spaced components or peaks, such as those of phages lambda and 2C. Once the CsCl gradient is well established, the two components' approaches to equilibrium will each be exponential (see Appendix). Their trapezoid-like pre-equilibrium profiles will appear, first superimposed, and then gradually narrow into the two equilibrium peaks. It is however unlikely that this simple problem of two well-separated, narrow bands can be satisfactorily generalized to the continuous wide band or GC distribution of a mammal. Rather than tinkering an approximate solution from off-the-shelf pieces, it would again seem indicated to begin afresh from first principles, where one has a firm grip on the assumptions one is making at every step of a derivation. The calculations of [35] treat the approach to equilibrium in the special case where dynamic changes of the DNA/CsCl solvation are negligible, and one possible path might begin from there.

Conclusion and Perspectives

A raw AUC profile of DNA in a CsCl density gradient, and its underlying GC distribution, change when a DNA sample is substituted by a sample from the same species but having a different molecular weight. The way in which the profile changes can report functionally relevant, statistical properties of the genome and its genes, a fact that renders CsCl profiles especially useful when a genome has not been sequenced.

Base compositional information can also help in reconstructing or confirming phylogenetic relationships (see [34] for a GC-based study of rodents). AUC-derived GC distributions can be phylogenetically informative, although satellite DNA contributions may need to be discounted: such highly repetitive DNA can band anomalously and, even when it does not, its rapid changes usually amount to noise except at the population or species levels.

Many genomes exhibiting substantial GC heterogeneity at the 50–100 kb level have recently been sequenced. Such sequences have amply confirmed earlier rigorous deductions from AUC [14, 24, 27, 36–38] and now point the way to refined post-processing of CsCl absorbance profiles. Where there are subtle but unexpected differences between AUC profiles and the GC histograms from scans of whole-genome sequences, they can indicate gaps in the sequence or gaps in our understanding of macromolecules' collective behavior in density gradients. CsCl profiles can be simulated or produced *in silico* from a genome sequence, incorporating facts and hypotheses to account for different factors that affect experimental profiles. Sequence-AUC comparisons can then be designed to fine-tune the hypotheses.

Acknowledgement We thank Gabriel Macaya for helpful discussions.

Appendix

Historical and Theoretical Details

Compositionally Homogeneous DNA. Almost all of the important quantitative physico-chemical and biophysical work on salt density gradient problems was done between 1957 and the mid-1970's, and progress in deriving appropriate formulae came to an almost complete halt when molecular cloning began.

Most early applications of CsCl density AUC involved DNA samples that were homogeneous in buoyant density and monodisperse in molecular weight, as is the case for preparations of intact complete phage DNAs (or nearly homogeneous and monodisperse, as for some bacterial DNAs), so that most of the early theoretical treatments also focussed on such DNA. The studies estimated, for example, molecular weight from band widths, i.e., from profile standard deviations, and calculated how various factors relevant to a routine AUC run would influence the position, shape and width of a band at sedimentation equilibrium. The factors included electric fields (DNA as a polyelectrolyte) [39], nonideality/solvation [40], virial effects/concentration dependence [19], pressure/compressibility [41, 42], methylation [43], other modifications of DNA [11], the presence of highly repetitive DNA [18, 44], and light bending [45].

Absorbance profile shapes depend primarily on the molecular weight, on the GC heterogeneity present, and on any anomalous banding behavior that may affect some but not the rest of the DNA. In the case of a homogeneous sample the molecular weight M is the most obvious factor. The expected profile is then roughly Gaussian, and the variance has the form

$$\sigma^2_{\text{total}} = a/M\,, \tag{3}$$

where a is a proportionality constant (see [40] for details).

Profiles' shapes also depend on the average concentration of DNA, i.e., on the amount of DNA loaded, all other things being equal. This concentration dependence was quantitatively analyzed by Schmid and Hearst [19, 46], who found that the Gaussian profiles of all phages tested had widths that increased exponentially with increasing concentration c, i.e. they had the form

$$\sigma_{\text{total}} = \sqrt{a/M}\,\mathrm{e}^{Bc}\,. \tag{4}$$

This behavior is what one would expect from an unusually strong virial effect. The widening of the profile is highly reproducible when the same phage is analyzed (Gabriel Macaya, personal communication describing unpublished data). On the other hand, some phages' profiles widen more rapidly with increasing concentration than others, i.e., B is species-specific.

How a homogeneous sample of macromolecules approaches its equilibrium distribution in density gradients

has been the topic of several articles. One detailed modelling effort [47] assumed some fixed functional forms *a priori* in order to extend a diffusion-free solution that had, in essence, been obtained earlier from the Lamm equation in another context via the method of characteristics [48]. The articles also included an elegant study for DNA in the special case of a pre-formed gradient [49]. The results derived in that study allow one to estimate a homogeneous sample's molecular weight by observing the banding of the DNA during the last hours of its approach to equilibrium: once the CsCl gradient has been established, the width and mean of a band's profile approach their equilibrium values exponentially [50]. A log-linear plot of the remaining difference versus time then yields a straight line, whether one is observing the standard deviation or the mean. In the former case, the molecular weight can be calculated from the straight line's slope. The calculation takes solvation (non-ideality) into account [50], and the results for phages agree favorably with molecular weights calculated via sedimentation velocity and/or now by whole-phage sequencing. The approach to equilibrium in density gradients is, interestingly, the topic also of two much more recent studies [35, 51], and a solution, for the limiting case of no interaction between DNA/CsCl and water, is now included in the program SEDFIT (http://www.analyticalultracentrifugation.com).

Compositionally Heterogeneous DNA. The conceptual simplicity of homogeneous samples, exemplified by the short genomes of bacteriophages, prompted some excellent theoretical work. Such short genomes are, however, rare among many of the species of current interest, and the larger chromosomes of prokaryotes and especially of eukaryotes are invariably broken into fragments during routine DNA extraction. In addition, genomes of eukaryotes such as deuterostome or protostome animals or angiosperm plants are characterized by marked intragenomic contrasts in GC. As a result, samples of total nuclear DNA from those genomes exhibit substantial intermolecular (inter-fragment) heterogeneity, so they are represented by wider main bands in CsCl at equilibrium.

Indeed, the original 1957 paper on CsCl gradient AUC [2] already showed absorbance profiles of a phage, and of calf, mentioning that in calf "the skewness in the resultant band indicates heterogeneity in [mean buoyant] density", and that such "density heterogeneity may be compositional or structural in origin".

A detailed discussion of heterogeneous DNA was given by Sueoka two years later [52]. We assume here for simplicity that we are analyzing a sample in which none of the DNA bands anomalously, i.e., the time-averaged position of each of the jittering DNA molecules is where Eqs. 1 and 2 predict. If some DNA does band anomalously, its GC must be replaced by its effective GC in the following, i.e., by the GC to which its buoyant density would correspond; the effective GC can even be negative in the case of long poly-A repeats [44, 53]. We also assume here that we are near the theoretical limit of infinite dilution, so that there are no virial or other concentration-dependent effects. The absorbance profile of the DNA is then a convolution of the true GC distribution, after converting to appropriate units, and a Gaussian point spread function (i.e., a kernel, or filter) that broadens the GC distribution via diffusion. The total variance σ^2_{total} of the band is therefore just the sum of the variance σ^2_{GC} due to true GC heterogeneity, plus the variance $\sigma^2_{\text{diffusion}}$ caused by the Brownian motion of molecules.

When molecular weight is constant (monodisperse sample, fixed-length fragments or molecules), the latter variance is inversely proportional to the molecules' common length or molecular weight M:

$$\sigma^2_{\text{total}} = \sigma^2_{\text{GC}} + a/M\,. \tag{5}$$

When molecular weight is not constant (polydisperse sample, variable-length fragments or molecules), the point spread function is in general not Gaussian. Indeed, our point spread function corresponds simply to the absorbance profile of a sample that is homogeneous in GC. If that sample is homogeneous but polydisperse, its absorbance profile is the weighted superposition of the profiles that would be obtained for each of the different molecular weights present. Even if our sample is heterogeneous and polydisperse, but exhibits no correlation between molecular weight and GC, a convolution can still be assumed. The resultant point spread function is then a weighted sum (or integral) of Gaussians, and a suitably modified version of Eq. 5 holds.

The molecular weight M of a DNA sample can be determined via an independent method such as sedimentation velocity or (more recently) pulsed-field gel electrophoresis. If one knows the proportionality factor a in Eq. 5, one can then quickly calculate the GC heterogeneity. The best estimate of this factor is that given by Schmid and Hearst [20], who included solvation (non-ideality) in their treatment, checked their formula using several phages of known lengths, and described its use for heterogeneous DNA. Since a is only weakly dependent on GC, since the GC levels of heterogeneous genomes' fragments are usually between 30% and 70%, and since other variables such as temperature remain standard for AUC runs, we hardly sacrifice any accuracy by treating a as a constant. For standard deviations in units of GC% we obtain for a, at 25° and other standard conditions, the value (44.5 ± 0.5) kb [14], which is at least as accurate as typical estimates of molecular weights from pulsed-field gels.

The above estimates and observations can be theoretically justified only at infinite dilution. This simplifying assumption is, however, unrealistic in practical AUC situations. AUC runs are not done at infinite dilution of the DNA, and not even at high dilutions: maximal absorbances (optical densities) should not be less than about 0.25. The

estimate of a discussed above is therefore a rough approximation, which we use only because at present no more accurate alternative is available (see below, and main text). The roughness of the approximation can depend on the genome of interest and/or on the part of the profile considered.

As Sueoka's early calculations [52] showed, the CsCl profile of calf thymus, i.e., of the bovine genome, was wider than those of certain bacteria and phages because of true heterogeneity among the (time-averaged) buoyant densities of the calf's DNA fragments, and not trivially because of any molecular weight differences among the samples being compared. It was not until 14 years later, however, that the CsCl profile of the cow genome could be quantitatively explained in terms of its genomic GC. Indeed, its heterogeneity comes in part from the GC heterogeneity of single-copy DNA, but also in part from the massive amounts (25%) of highly repetitive satellite DNA that are present in the cow genome [16]. Such satellite DNA can be cryptic, and can easily broaden a profile, especially when it bands anomalously. Other examples in point are the guinea pig profile [18] and the mouse profile that is still included in standard biology textbooks to illustrate the concept of satellites.

In summary, at infinite dilution a solution for a homogeneous sample can be transferred to the case of a heterogeneous sample: the solution of the homogeneous problem becomes the point spread function for the heterogeneous problem. The convolution principle that allows this guarantees that the variances of the spreading and of the GC distribution will add up to give (in units of equivalent GC%) the total profile variance.

Deconvolving a CsCl profile might seem the most direct way to obtain or extract a GC distribution from an absorbance profile. In view of the current resolution limits of commercially available analytical ultracentrifuges, however, it is both numerically easier and more instructive to work in the other direction. For example, our experience has shown that a truncated exponential distribution convolved with a Gaussian is typically a very good fitting function for GC distributions or (with a wider Gaussian, to accommodate diffusion) raw CsCl profiles of a mammal's total DNA. Satellites are then often visible as bumps that locally deviate from the best fit [14, 34]. Similarly, when a genome sequence is available in its entirety (not yet common, even when the sequence is declared "finished"), its GC distribution can be convolved with model spread functions and the resulting distribution can then be compared with experimentally obtained CsCl profiles. Partial sequences allow rough comparisions.

In a convolution/deconvolution problem, the spreading or broadening must be the same at all parts of the GC distribution or profile. A natural generalization of such a problem is a corresponding folding/unfolding problem [54], where this requirement is relaxed. In a folding model, the spread function and its width (extent of broadening) could be allowed to depend directly on the GC (in our case), and therefore indirectly on any other variables that are unambiguously specified by the GC. Folding models might therefore be useful in some situations where convolution models are too restrictive, although they can unfortunately not be applied to solve two problems for heterogeneous DNA discussed in this paper, concentration dependence and the approach to equilibrium. In profiles of heterogeneous DNA, aggregation and possibly other effects can counteract the flattening effect of concentration that is observed for species with homogeneous DNA such as bacteriophages. Thus, the homogeneous case cannot be generalized via folding to yield the concentration dependence for a heterogeneous species such as human. Similarly, no stage in the approach to equilibrium admits a folding model: GC-rich and GC-poor DNA molecules are not much quicker or slower in separating from each other than in approaching equilibrium.

References

1. Bernardi G (2004) Structural and evolutionary genomics: Natural selection in genome evolution. Elsevier, Amsterdam etc.
2. Meselson M, Stahl FW, Vinograd J (1957) Proc Natl Acad Sci USA 43:581
3. Holmes FL (2001) Meselson, Stahl, and the replication of DNA. Yale University Press, New Haven London
4. Sueoka N, Marmur J, Doty P (1959) Nature 183:1429
5. Rolfe R, Meselson M (1959) Proc Natl Acad Sci USA 45:1039
6. Marmur J, Doty P (1959) Nature 183:1427
7. MacGregor IK, Anderson AL, Laue TM (2004) Biophys Chem 108:165
8. Collins FS, Green ED, Guttmacher AE, Guyer MS (2003) Nature 422:835
9. Powledge TM (2004) Genome Biology (Research News) 18 Nov 04
10. Smith C (2005) Nature 435:991
11. Schildkraut CL, Marmur J, Doty P (1962) J Mol Biol 4:430
12. Ifft JB, Voet DM, Vinograd J (1961) J Phys Chem 65:1138
13. Thiery JP, Macaya G, Bernardi G (1976) J Mol Biol 108:219
14. Clay O, Douady CJ, Carels N, Hughes S, Bucciarelli G, Bernardi G (2003) Eur Biophys J 32:418
15. van Holde KE, Johnson WC, Ho PS (1998) Principles of physical biochemistry. Prentice-Hall, Upper Saddle River NJ
16. Filipski J, Thiery JP, Bernardi G (1973) J Mol Biol 80:177
17. Mazrimas JA, Hatch FT (1972) Nature New Biol 240:102
18. Corneo G, Ginelli E, Soave C, Bernardi G (1968) Biochemistry 7:4373
19. Schmid CW, Hearst JE (1969) J Mol Biol 44:143
20. Schmid CW, Hearst JE (1972) Biopolymers 11:1913
21. Macaya G, Thiery JP, Bernardi G (1976) J Mol Biol 108:237

22. Hudson AP, Cuny G, Cortadas J, Haschemeyer AE, Bernardi G (1980) Eur J Biochem 112:203
23. Cuny G, Soriano P, Macaya G, Bernardi G (1981) Eur J Biochem 115:227
24. Pavlíček A, Pačes J, Clay O, Bernardi G (2002) FEBS Lett 511:165
25. Beran J (1994) Statistics for long-memory processes. Chapman & Hall/CRC, Boca Raton etc.
26. Zivot E, Wang J (2003) Modeling financial time series with S-PLUS. Springer-Verlag, New York NY
27. Clay O, Carels N, Douady C, Macaya G, Bernardi G (2001) Gene 276:15
28. Clay O (2001) Gene 276:33
29. Li W, Holste D (2005) Phys Rev E 71:041910
30. Bucciarelli G, Bernardi G, Bernardi G (2002) Gene 295:153
31. Bernardi G, Bernardi G (1986) J Mol Evol 24:1
32. Jabbari K, Bernardi G (2004) Gene 333:183
33. Kulathinal RJ, Bettencourt BR, Hartl DL (2004) Science 306:1553
34. Douady C, Carels N, Clay O, Catzeflis F, Bernardi G (2000) Mol Phylogenet Evol 17:219
35. Schuck P (2004) Biophys Chem 108:187
36. Lander ES, Linton LM, Birren B et al. (2001) Nature 409:860
37. Bernardi G (2001) Gene 276:3
38. Bernaola-Galván P, Oliver JL, Carpena P, Clay O, Bernardi G (2004) Gene 333:121
39. Yeandle S (1959) Proc Natl Acad Sci USA 45:184
40. Schmid CW, Hearst JE (1971) Biopolymers 10:1901
41. Vinograd J, Hearst JE (1962) Fortschritte der Chemie organischer Naturstoffe 20:372
42. Szybalski W (1968) Methods Enzymol 12:330
43. Kirk JT (1967) J Mol Biol 28:171
44. Wells RD, Blair JE (1967) J Mol Biol 27:273
45. Hearst JE, Vinograd J (1961) J Phys Chem 65:1069
46. Hearst JE, Schmid CW (1973) Methods Enzymol 27:111
47. Dishon M, Weiss GH, Yphantis DA (1971) Biopolymers 10:2095
48. Fujita H (1956) J Am Chem Soc 78:3598
49. Hearst JE (1965) Biopolymers 3:1
50. Schmid CW, Hearst JE (1972) Biopolymers 11:1765
51. Minton AP (1992) Biophys Chem 42:13
52. Sueoka N (1959) Proc Natl Acad Sci USA 45:1480
53. Sober HA (ed) (1968) Handbook of biochemistry: Selected data for molecular biology. CRC, Cleveland Ohio
54. Roe BP (1992) Probability and statistics in experimental physics. Springer-Verlag, New York etc.