Eur Biophys J (2003) 32: 418–426
DOI 10.1007/s00249-003-0294-y

ARTICLE

**Oliver Clay · Christophe J. Douady
Nicolas Carels · Sandrine Hughes
Giuseppe Bucciarelli · Giorgio Bernardi**

Using analytical ultracentrifugation to study compositional variation in vertebrate genomes

Received: 11 June 2002 / Revised: 19 December 2002 / Accepted: 28 January 2003 / Published online: 9 April 2003

Abstract Although much attention has recently been directed to analytical ultracentrifugation (AUC), the revival of interest has hardly addressed the applications of this technology in genome analysis, and the extent to which AUC studies can quickly and effectively complement modern sequence-based analyses of genomes, e.g. by anticipating, extending or checking results that can be obtained by cloning and sequencing. In particular, AUC yields a quick overview of the base compositional structure of a species' genome even if no DNA sequences are available and the species is unlikely to be sequenced in the near future. The link between AUC and DNA sequences dates back to 1959, when a precise linear relation was discovered between the GC (guanine + cytosine) level of DNA fragments and their buoyant density in CsCl as measured at sedimentation equilibrium. A 24-hour AUC run of a high molecular weight sample of a species' total DNA already yields the GC distribution of its genome. AUC methods based on this principle remain sensitive tools in the age of genomics, and can now be fine-tuned by comparing CsCl absorbance profiles with the corresponding sequence histograms. The CsCl profiles of vertebrates allow insight into structural and functional properties that correlate with base composition, and their changes during vertebrate evolution can be monitored by comparing CsCl profiles of different taxa. Such comparisons also allow consistency checks of phylogenetic hypotheses at different taxonomic levels. We here discuss some of the information that can be deduced from CsCl profiles, with emphasis on mammalian DNAs.

Electronic Supplementary Material Supplementary material is available for this article if you access the article at http://dx.doi.org/10.1007/s00249-003-0294-y. A link in the frame on the left on that page takes you directly to the supplementary material.

Keywords Analytical ultracentrifugation · Base composition · Density gradients · Isochores · Phylogeny

Abbreviations AUC: analytical ultracentrifug (ation) · bp, kb, Mb: base pairs, kilobase pairs (10^3), megabase pairs (10^6) · GC: molar fraction of guanine and cytosine in DNA, guanine-cytosine base pair · GC_3: GC in third codon positions of coding genes · H: compositional (GC) heterogeneity · l: segment/molecule/fragment length · M: molecular weight, molar mass · OD: optical density, absorbance · PSF: point spread function · ρ: buoyant density · σ: standard deviation · vWF: gene for von Willebrand factor · XL-A: Beckman analytical ultracentrifuge with absorbance optics

Presented at the Conference for Advances in Analytical Ultracentrifugation an Hydrodynamics, 8–11 June 2002, Grenoble, France

Electronic Supplementary Material Supplementary material is available for this article if you access the article at http://dx.doi.org/10.1007/s00249-003-0294-y. A link in the frame on the left on that page takes you directly to the supplementary material.

O. Clay · C. J. Douady · N. Carels · S. Hughes
G. Bucciarelli · G. Bernardi (✉)
Laboratory of Molecular Evolution,
Stazione Zoologica Anton Dohrn,
Villa Comunale, 80121 Napoli, Italy
E-mail: bernardi@alpha.szn.it
Fax: +39-081-2455807

C. J. Douady
Medical Biology Centre, Biology and Biochemistry,
Queen's University, 97 Lisburn Road,
Belfast, BT9 7BL, UK

S. Hughes
Laboratoire de Biométrie et Biologie Evolutive, UMR 5558,
Université Claude Bernard Lyon 1, Villeurbanne, France

Present address: C. J. Douady
Department of Biochemistry and Molecular Biology,
Faculty of Medicine, Dalhousie University,
B3H 4H7 Halifax, Nova Scotia, Canada

Present address: N. Carels
Centro de Astrobiologia, INTA edificio S-18,
Ctra de Torrejon a Ajalvir,
28850 Torrejon de Ardoz, Spain

Introduction

The aim of this article is to provide a bird's eye view of some simple uses of CsCl gradient ultracentrifugation for studying GC variation within and among vertebrate genomes, in order to illustrate its potential for genomics and for evolutionary studies.

GC is defined as the molar fraction of guanine and cytosine in a molecule or segment of DNA, i.e., the proportion of its base pairs that are GC rather than AT. This most fundamental base compositional property of double-stranded DNA can be easily measured in an analytical ultracentrifuge (or in a preparative ultracentrifuge followed by high-resolution fractionation). The measurements are made in density gradients of heavy salts. Of these salts, cesium chloride is the most widely used. It is commercially available in optical-grade quality, it allows a faithful (linear) portrayal of GC distributions in an analytical ultracentrifuge (AUC), and it permits high-resolution fractionation according to GC content in a preparative ultracentrifuge.

The technique of density gradient ultracentrifugation was introduced in 1957 by Meselson, Stahl and Vinograd (1957). The principle is simple: a heavy salt of low molecular weight in solution will, upon centrifugation, establish a density gradient. At sedimentation equilibrium, double-stranded DNA molecules having a given GC will be found neither at the meniscus nor in the pellet, but in a narrow band within the density gradient. One therefore places the DNA together with the salt solution in the ultracentrifuge cell and allows the salt and DNA to reach equilibrium, which under standard conditions is attained within 24 hours. The GC level(s) of the DNA can be read from its position in the solution.

Soon after the first experiments, it was discovered (Marmur and Doty 1959; Rolfe and Meselson 1959; Sueoka et al. 1959; Schildkraut et al. 1962) that in CsCl gradients the GC level of a double-stranded DNA molecule exhibits a remarkably linear relationship to the position of the molecule at sedimentation equilibrium (this is not the case in cesium sulfate gradients, for example, where the relation is curved; see Szybalski 1968). More precisely, the GC level of the DNA molecule is linearly related to the density of the CsCl solution at its equilibrium position. This density is called the buoyant density of the DNA in CsCl and is, in turn, essentially a linear function of the radial distance from the ultracentrifuge axis. One can therefore measure not only the GC level of a sample of compositionally similar molecules, but also the GC distribution of compositionally heterogeneous genomes such as the human genome, which spans a GC range from just under 30% GC to just over 60% GC (at scales up to several megabases). Indeed, the CsCl absorbance profile of high molecular weight DNA fragments is, after a linear transformation of the horizontal axis, the GC distribution of the fragments, to a very good approximation. Only when fragments are shorter than about 15 kb (i.e., 10×10^6 Daltons) does diffusion seriously distort the profile. Similarly, only when DNA fragments are heavily methylated or otherwise modified (as in T-even phages), highly repetitive, or denatured do they shift from their expected equilibrium positions.

The CsCl method has been of central importance in understanding compositional variation along mammalian chromosomes; some of the main conclusions were drawn well before any DNA sequences were known (Filipski et al. 1973; Macaya et al. 1976; Thiery et al. 1976). An early result was the discovery that mammalian genomes are organized into long, compositionally fairly homogeneous regions, called isochores. By comparing absorbance profiles of the same species for different fragment sizes (molecular weights), and by monitoring the profiles' resistance to narrowing as the fragment sizes are decreased, one can infer statistical properties of the mosaic GC variation along its chromosomes (Macaya et al. 1976; Cuny et al. 1981; for comments on the concordance with recent draft sequence results, see Bernardi 2001; Clay 2001; Clay et al. 2001).

When the first genes were sequenced in the mid-1980s, hybridization studies on compositional fractions showed that the GC levels in the three codon positions of the genes were correlated to the GC level of the DNA surrounding the genes (Bernardi et al. 1985; Bernardi and Bernardi 1986), most notably for the third codon positions (the correlation coefficient lies between 0.65 and 0.85, depending on the reliability of the gene database and the presence of repetitive DNA).

The GC levels of third codon positions can be changed while leaving the encoded amino acid sequence essentially intact, so that such a freedom would allow these third positions to fulfil a role at the DNA level. The correlations between genic and intergenic DNA were therefore interpreted as reflecting compositional constraints, acting on both coding DNA (which comprises only 3% of the DNA in the human genome, for example) and on noncoding DNA (which comprises the rest of the genome). The correlations allowed quantitative, genome-wide information to be deduced for the 30,000–70,000 genes in a mammalian nuclear genome, on the basis of ultracentrifugation experiments and relatively few sequenced genes. An example is the gene distribution of the human genome, which was calculated in 1991 using CsCl analysis and just over 1000 coding sequences (Mouchiroud et al. 1991). It has now been quantitatively confirmed by the draft genome sequence (IHGSC 2001; see Bernardi 2001 for a discussion of this concordance). The important finding was that the number of genes per megabase in the GC-richest regions of the human genome (which, as can be seen from absorbance profiles, contain only a small percent of the DNA in the genome) is up to 20 times higher than in the GC-poorest regions, which are practically gene deserts. This finding, and the differences in GC heterogeneity between primates and fishes (see below), are rigorous quantitative deductions. Such findings are not merely hypotheses or indications that were "suggested"

by "physical methods", as some genome sequencers or annotators imply. If that were true, the new sequence results in question (IHGSC 2001; Aparicio et al. 2002) would indeed constitute a first rigorous proof, or even a discovery. Instead, such sequence results are an interesting and useful confirmation. The accuracy of quantitative conclusions from ultracentrifuge work is apparently still widely underestimated, but the precise agreement between recent sequence results and much earlier ultracentrifuge results will hopefully revive a general awareness of the power of the CsCl/AUC method, and an interest in the biological information it can yield.

Within the genomes and chromosomes of vertebrates, GC co-varies with several well-documented structural and functional features of the DNA and of the genes it contains. Some of these correlates of regional or isochore GC level, which include replication timing, gene density, methylation and CpG island density, and chromosomal territories in interphase, are reviewed and described in Bernardi (2000a, 2001), Tenzen et al. (1997) and Saccone et al. (2002).

Although the nuclear genomes of human, mouse and pufferfish (as well as those of fly, a nematode, yeast, rice, a dicot, and more than 60 prokaryotes) have now been entirely or largely sequenced, the majority of vertebrate orders are still represented by less than a dozen sequenced nuclear genes. For understanding the relations between their genomes, and the base compositional changes that have occurred during their evolution, the CsCl methodology remains a valuable tool. Not many methods exist today that are as efficient at informing about an uncharacterized vertebrate genome in 24 hours, or at extending the information in a few genic sequences to obtain results of genome-wide validity.

Materials and methods

Standard conditions

We will denote by "standard conditions" the following situation. The temperature should be 25 °C. The DNA should consist of double-stranded fragments that are all above 2–3 kb and contain low percentages of tandem highly repetitive (satellite) DNA. Less than 3% of the bases should be methylated or otherwise modified, a condition that is fulfilled for vertebrates (Jabbari et al. 1997) but not for some plants, and not for T-even phages, for example, where corrections are needed (Schildkraut et al. 1962; Kirk 1967). Standard speeds are 44,000 rev/min for CsCl work using the Beckman XL-A analytical ultracentrifuge (44,770 rev/min for the model E, 35,000 rev/min for preparative ultracentrifugation followed by fractionation; De Sario et al. 1995); the standard wavelength is 260 nm. Concentrations of DNA should result in maximal absorbance (i.e., optical density or OD) between 0.3 and 1.0. A time of 24 h should be allowed for sedimentation equilibrium to be reached. Bacteriophage 2c (which has modified bases that raise its buoyant density to 1.7420 g cm^{-3}) can be used as a marker. Minor adjustments of the average CsCl density chosen for the solution, in order to guarantee optimal or comparable band positions in the gradient ("isopycnic" point; see e.g. Hearst et al. 1961), are not discussed here.

Converting radial position to GC level

The GC level of a DNA molecule or fragment is calculated from its expected buoyant density ρ in CsCl (Schildkraut et al. 1962): $GC = 100\% \times (\rho - 1.660 \text{ g cm}^{-3})/0.098$. The buoyant density ρ is, in turn, calculated from the fragment's mean distance r from the axis of the ultracentrifuge at sedimentation equilibrium (Ifft et al. 1961): $\rho = \rho_m + \omega^2(r^2 - r_m^2)/(2\beta_B)$. Here, ρ_m and r_m are the buoyant density and radial position of a suitable marker DNA (such as bacteriophage 2c), ω is the angular speed and β_B is $1.190–1.195 \times 10^9$ (cgs units) for Beckman models E and XL-A under standard conditions (cf. Thiery et al. 1976). Since the distances between banding DNA molecules are very small compared to the distances from the axis, $r^2 - r_m^2 \approx 2r_m(r - r_m)$, i.e. both of the above equations, and therefore the relation between GC and position, are essentially linear.

Comments on molecular weight

High molecular weight (molar mass) DNA contains more information than low molecular weight DNA. Indeed, a long sequence contains all the information about its subsequences, but not vice versa: reassembling a long sequence from its fragments is a non-trivial, and sometimes intractable, task (as the persisting gaps in the human draft sequence attest). Furthermore, low molecular weight DNA is subject to strong diffusion. The diffusion washes out details of such DNAs' GC distribution, such as subtle bumps, that cannot be recovered by a deconvolution in practice.

Routine extraction always fragments DNA. However, proper precautions allow one to obtain long fragments (high molecular weights). Depending on the protocol or kit used, and possibly on the endonuclease activity of the sampled species, one may then expect fragment lengths from about 20 kb up to several hundred kb. (In the following, molecular weights, i.e. molar masses, will be expressed in kilobases, which are linearly related to Daltons: 3 kb corresponds to 2×10^6 Da, 300 kb to 200×10^6 Da). Appropriate precautions must be observed during the gathering and preservation of tissues (Dessauer et al. 1996), and during extraction and injection of the DNA macromolecules into the cell of the analytical ultracentrifuge. Subfragments can be obtained via mechanical shearing, e.g. by passing the DNA repeatedly through a syringe (Macaya et al. 1976). To ensure precision, extracted DNA may be further purified by chromatography on hydroxyapatite in the presence of 3 M KCl (Filipski et al. 1973; Thiery et al. 1976), although this additional step does not appear necessary for bulk genome screening of high molecular weight DNAs, i.e. for the comparison of their GC distribution parameters as described in this article.

When very long fragments (~100 kb or longer) are injected into the ultracentrifuge cell through its usual narrow opening, they will be sheared. This injection step can be avoided by placing the DNA directly inside the cell and only then reassembling the cell around the DNA (see Macaya et al. 1976 for details).

Theory

Given a CsCl absorbance profile of a sample of vertebrate DNA, we face two tasks. The first task is to accurately estimate the GC distribution from the fragments in the sample, and this is discussed in this section. The rudiments of CsCl gradient ultracentrifugation are introduced in van Holde et al. (1998, sections 5.2 and 13.2), Berg (1983) and Fujita (1962), and will not be reviewed here.

The second task is to interpret the GC distribution thus obtained. It involves comparing it with GC distributions for the same species at other molecular weights, or with GC distributions for other species obtained at similar molecular weights, and interpreting the differences in terms of genome structure, function and evolution. Some aspects of this second task will be presented in the Results and discussion.

Samples of very high molecular weight (> 50 kb)

The DNA of vertebrates is characterized by a persistent base compositional heterogeneity: for fragment lengths up to several hundred kilobase pairs, the GC distribution remains wide. In contrast to bacterial genomes, for example, which are homogeneous at such scales (partly as a result of their much smaller size), the range of GC levels within many mammalian genomes extends over more than 30% GC. This means that the absorbance profiles of vertebrates, obtained in CsCl, are much wider than those of bacteria or bacteriophages, which were the prime objects of study during the early years of CsCl density ultracentrifugation. As a result, many of the effects that were studied in detail at that time, including diffusion and light bending, contribute only marginally to the shape or width of high molecular weight DNA from mammalian and other vertebrate genomes.

50 kb is an approximate limit for fragment lengths, above which GC distributions of mammalian genomes maintain essentially the same width, up to lengths approaching or exceeding 1 Mb (see Results and discussion). This means that, for such samples, molecular weight (and its possible polydispersity) does not influence the profile width, a fact that has implications for the DNA sequences of mammalian chromosomes. 50 kb is also an estimate of the limit above which one can safely neglect the diffusion broadening of vertebrate profiles.

Monodisperse samples of intermediate molecular weight (15–50 kb)

DNA molecules or fragments will diffuse about their expected positions in the density gradient, even after sedimentation equilibrium has been reached. Short molecules will usually wander farther than long molecules, although the average number of molecules at any position will not change, i.e. there will be no more net motion at a macrosopic scale. If the fragments or molecules are shorter than about 50 kb, the distances over which they diffuse may become appreciable, compared to the width of the vertebrate's CsCl profile. Unless one properly discounts the resulting diffusion contribution to the absorbance profile, one may mistakenly interpret such diffusion broadening as true GC heterogeneity. It is therefore important to know how to recover the GC distribution from absorbance profiles of samples that have intermediate molecular weights (15–50 kb).

It is the hydrated (solvated) macromolecules, not the pure DNA molecules, that experience diffusion. In the (rare) case in which there is excessive aggregation, the units on which diffusion acts are the aggregates. For DNA and/or salt solutions, the conditions for aggregation are apparently still not yet well understood (Samal and Geckeler 2001). In some cases, aggregation of DNA in CsCl solutions may be caused by highly repetitive DNA sequences, or by the presence of DNA molecules with "sticky ends" (Macaya et al. 1976). In the following, we assume that aggregation is negligible.

We first treat the simplest case of a monodisperse sample, i.e. when we know that all fragments have the same length. An absorbance profile of monodisperse DNA in a CsCl density gradient can be viewed as the convolution of the GC distribution of the DNA and a Gaussian "point spread function" (PSF) that describes the diffusion of the solvated DNA fragments around their expected equilibrium positions in the gradient. For this picture to be valid, double-stranded DNA molecules of identical length but different GC contents should experience essentially the same diffusion. Experience and theory indicate that this assumption is correct, to a good approximation. When the diffusion is essentially independent of GC, the GC distribution can be extracted from the absorbance profile by a deconvolution, if the diffusion PSF is known (if the diffusion were strongly GC dependent, unfolding would be needed instead of deconvolution). Since a Gaussian PSF is uniquely determined by its variance (the mean of a PSF is always 0), that variance is what we must now determine.

If we just need to calculate the heterogeneity in GC, and not the full GC distribution, then, equipped with the variance of the diffusion PSF, we can use the additivity of variances that is valid for convolutions (Sueoka 1959):

$$\sigma^2_{\text{total}} = H^2 + \sigma^2_{\text{diffusion}} \tag{1}$$

and we do not need to explicitly deconvolve. In other words, to obtain the variance in GC we simply subtract the diffusion variance from the profile variance. Here, we have used or converted to GC% throughout; the standard deviation of the GC distribution, denoted by H, is called the *compositional heterogeneity* of the DNA, as in previous work.

If all other conditions are kept constant, the diffusion variance $\sigma^2_{\text{diffusion}}$ is assumed to depend only on the length of the DNA fragments. This variance could therefore be determined experimentally: for a sample of DNA in which all fragments have identical GC content, as well as identical length l, we have $H = 0$.

The total CsCl profile would then be the diffusion profile or PSF (valid also for other DNAs of the same length), and its standard deviation would be $\sigma_{\text{diffusion}}$. In astronomy, such an experimental procedure corresponds (as the name "point spread function" suggests) to finding the PSF or "broadening" contributed by a local atmosphere + telescope system, by measuring the image of a distant star.

Rather than pursing a purely empirical approach, we use the calculations of Schmid and Hearst (1972). The diffusion contribution to a CsCl profile's variance, expressed in equivalent GC%, is then expected to be inversely proportional to the fragments' or molecules' length l in kb, and only very mildly dependent on GC via the buoyant density ρ:

$$\sigma^2_{\text{diffusion}} = \left(\frac{100\%}{0.098}\right)^2 \frac{\rho RT}{\beta_B^2 G M_{Cs}} \frac{1}{1000 l} \tag{2}$$

Here, ρ is the buoyant density of the profile at the approximate position of interest,$G = (1 + \Gamma')/\beta_{\text{eff}}$ is a buoyancy factor, which at 25 °C has been estimated at 7.87×10^{-10} cgs units (Schmid and Hearst 1969, 1971, 1972) and β_B is 1.195×10^{-10} cgs units under standard conditions (see above). Finally, M_{Cs} denotes the molecular weight per base pair of dry cesium DNA, 882 (M_{Cs} is 4/3 times the value for dry *sodium* DNA, i.e. 4/3 times the standard factor used for converting between Daltons and base pairs).

At 25 °C we then obtain, for example, in the range 30–70% GC:

$$\frac{44.0\text{ kb}}{l} < \sigma^2_{\text{diffusion}} < \frac{45.0\text{ kb}}{l} \tag{3}$$

where the limits correspond to 30% and 70% GC. Thus, for 1 kb fragments, diffusion will broaden the profile by the equivalent of 6.63–6.71% GC; for 44.5 kb fragments the broadening will be close to 1% GC. We see that the GC dependence is extremely small, even in wide mammalian profiles: if, in Eq. (3), we replace all GC values of interest by 50% GC, we incur an error in the diffusion spreading (PSF width) of less than 0.04% GC, for fragments longer than 1 kb.

The numerical values given in Eq. (3) lead to reasonable results for the vertebrate DNA samples we have tested. For example, the compositional heterogeneity H of fish genomes (which are the most compositionally homogeneous vertebrate genomes) approach, but never reach or fall below, those of random sequences, even for samples of quite low molecular weights (Bernardi and Bernardi 1990). The values in Eq. (3) should, however, still be considered approximate, for two reasons. The first reason is that the calculations of Schmid and Hearst (1972) apply after extrapolation to infinite dilution, and real concentrations of DNA under standard conditions (maximal absorbances between 0.3 and 1.0) will give slightly different values. The second reason is that the value for G may not be exactly constant for all DNAs. It was originally tested using two bacteriophages that have now been sequenced and that confirm a quite good accuracy of the G value given in 1972, but it may still marginally depend on sequence properties of the DNA other than its length or GC level.

In summary, the higher the molecular weight of the DNA, the less the correction will be, and the less one will need to rely on its accuracy.

Polydisperse samples of intermediate molecular weight (15–50 kb)

For most vertebrate samples of moderately high molecular weight, both the GC distribution and the diffusion vary only slightly over the range of molecular weights in the sample. Thus, the observed absorbance profile becomes very similar to that of a monodisperse sample in which all DNA fragments have a single length l_{eff}.

For pulsed-field gel (PFGE) estimates of the molecular weight distribution, the modal value of l is often chosen as an estimate of this length, partly because its measurement is straightforward. For sedimentation velocity estimates, the weight average is calculated from the sedimentation coefficient value $s^0_{20,\mathrm{w}}$ (Eigner and Doty 1965; Macaya et al. 1976; Thiery et al. 1976; Bernardi and Bernardi 1990). Since the diffusion variance is proportional to $1/l$, another choice of l_{eff} could be $1/\langle 1/l \rangle$, which is less than the mean $\langle l \rangle$ (Jensen's inequality). All these choices of l_{eff} give very similar values, even for moderate molecular weights. For example, a pulsed-field gel of a sample from the carnivorous marsupial *Murexia* yielded an almost perfectly Gaussian length distribution, with mean = mode = 28.2 kb and standard deviation 8.7 kb; $1/\langle 1/l \rangle$ was 24.3 kb, which remains close to the mode.

Other details

Further details and references are given in the review by Hearst and Schmid (1973). Some technical points related to CsCl gradient ultracentrifugation, inferring GC distributions, and fitting truncated exponential curves (a form that is largely conserved among mammals and many reptiles) are elaborated in the Supplementary material.

Results and discussion

Congruence between AUC and sequence data

Figure 1 shows a high molecular weight absorbance profile of total human DNA in a CsCl density gradient. It is the raw data from a randomly chosen, single scan of an XL-A analytical ultracentrifuge, with the horizontal axis calibrated in GC% instead of radial distance from the ultracentrifuge axis. This recalibration was done using only equations that have been in use since 1962 (see Materials and methods), and that have not been adjusted or reparametrized since then. The DNA was extracted from blood using a standard kit (Talent), and added to the CsCl solution with no special further purification. It can be seen that the absorbance profile, which is simply the GC distribution of the DNA fragments in the sample, agrees remarkably well with the corresponding GC distributions obtained from the draft sequence of the human genome, which became available 39 years after the equations were published (IHGSC 2001). The minor differences could be easily explained by fluctuations that are within the (visible) measurement error of the XL-A scan for the experimental curve, or by the tenth of the human DNA that is still missing from the draft sequence, i.e. the gaps. The only exception is a small amount of apparently very GC-poor DNA in the absorbance profile, which was not observed for other human or primate samples and is likely to represent a contaminant or data acquisition noise. The overall, positively asymmetric shape shown by the human GC distribution is found in most other mammals and, with some large variations in width, in almost all vertebrates.

The two sequence-derived GC distributions in Fig. 1 illustrate a remarkable property of mammalian genomes: the two curves are almost identical, yet they represent samples with (constant) fragment lengths of 100 kb and 300 kb, i.e. molecular weights differing by a factor 3. There is visible resistance to the gradual narrowing that one would intuitively expect from the central limit theorem, as fragment lengths increase. In particular, if the human genome sequence were com-

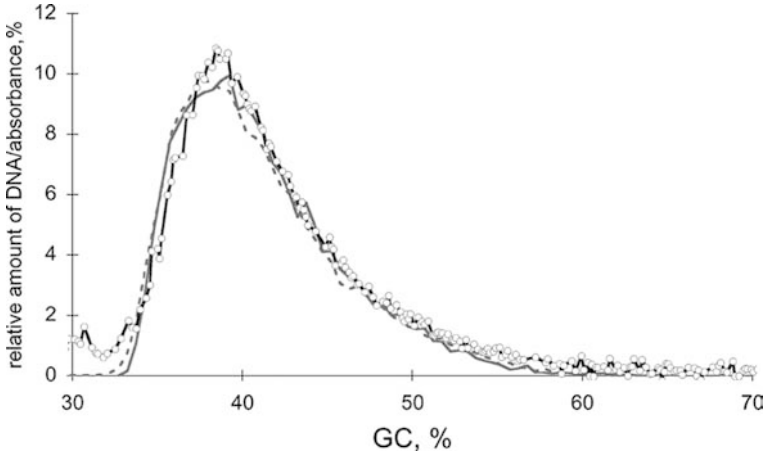

Fig. 1 Agreement between CsCl/AUC absorbance profiles of total human DNA and GC distributions of the draft genome sequence (IHGSC 2001). The absorbance profile (*circles connected by black lines*) is the raw output from a single XL-A scan of a sample of high (> 70 kb) molecular weight human DNA, with the horizontal axis recalibrated in % GC (see Materials and methods). The GC distributions from the human draft sequence, reproduced from Pavlíček et al. (2002a), are for fragment sizes of 100 kb (*solid grey curve*) and 300 kb (*dashed grey curve*). All distributions have very little DNA above 50–55% GC; however, the number of genes in these GC-richest regions is even slightly higher than in the much more abundant GC-poor regions. As a result, the gene density in the GC-richest regions of the human genome is 15–20 times higher than in the GC-poorest regions

posed of statistically independent (i.e., uncorrelated) and identically distributed base pairs, the GC distribution would already have become almost infinitely narrow (with a standard deviation $< 0.1\%$ GC), when fragment lengths reach $l = 100{,}000$ or 300,000 bp. Indeed, for such a sequence the binomial distribution would prescribe $\sigma^2 = \mu(1-\mu)/l$, where μ and σ are the mean and standard deviation of the GC distribution, expressed in units of GC%/100%. In contrast to this textbook prediction, it can be seen that both human GC distributions still extend over a range of at least 30% GC. The only possible explanation of this phenomenon, which was first observed for cow, mouse and other mammals in 1973–1976 using analytical ultracentrifugation (Filipski et al. 1973; Macaya et al. 1976), is that there is a strikingly nonrandom organization of GC level along the chromosomes of mammals, which persists at least up to scales of hundreds of kilobase pairs, and which is characterized by long-range autocorrelations and large-scale mosaicism. This point is of practical as well as biological importance: when the fragments in a mammalian DNA sample are all long (above about 50 kb), molecular weight polydispersity ceases to be an important problem. A monodisperse sample of 100 kb, a monodisperse sample of 300 kb and a polydisperse sample with fragment lengths ranging from 50 kb to 1 Mb (an unusually large polydispersity for routine extractions of DNA) will all yield absorbance profiles or GC distributions of practically identical width. Furthermore, in all three cases the profiles and GC distributions will give very similar information about the large-scale mosaic organization of GC level along the species' chromosomes.

The biological importance of the persistently wide mammalian profiles stems from the mosaic or isochore organization of chromosomes, to which this width corresponds. Isochores were originally defined as regions that are fairly homogeneous in GC and usually extend over expanses $> > 300$ kb (Macaya et al. 1976; Cuny et al. 1981). Isochore maps of the human genome (Li 2001; Oliver et al. 2001; Pavlíček et al. 2002a) now confirm that this definition can be kept, and show that some isochores even extend over many megabases, especially in the case of the GC-poorer isochores, which account for most of a genome's DNA (see Fig. 1).

The lower bound on the average size of isochores, $> > 300$ kb or $> > 200 \times 10^6$ Da, described the highest molecular weight sample that could be prepared and ultracentrifuged (without aggregation) in the 1970s using standard methods. For all preparations of mouse DNA analyzed in 1976, up to this molecular weight, the inferred GC distribution's width (compositional heterogeneity) remained essential unchanged, as is shown in Fig. 2. It could therefore be inferred that this behavior must apply also for much longer regions (see Pavlíček et al. 2002a for a discussion); size estimates of individual isochores became possible with new techniques (Bettecken et al. 1992; De Sario et al. 1996).

Figure 2 also shows a plot of the compositional heterogeneity obtained from the human draft sequence,

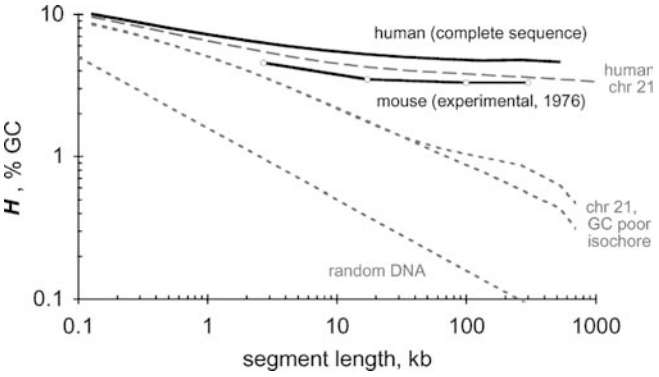

Fig. 2 Agreement between the behavior of GC distributions obtained from CsCsCl/AUC absorbance profiles and from genomic sequences, at different molecular weights. The ordinate shows compositional heterogeneity *H*, defined as the standard deviation of the GC distribution. Double-logarithmic plots are shown for the human draft sequence (*bold black curve*) and for mouse absorbance profiles at four molecular weights (*circles connected by bold black lines*), obtained using a model E AUC in 1976 (Macaya et al. 1976, fig. 8). Expected standard deviations in the absence of GC-GC autocorrelations are shown by the *grey dashed line* at the bottom (statistically independent and identically distributed, or "random", base pairs with $30\% < GC < 70\%$). For comparison, human draft sequence results are shown also for chromosome 21 and for the two halves of a long, ~7 Mb, GC-poor isochore in chromosome 21. Sequence plots are from Clay et al. (2001), where further details can be found

which has a very similar shape to that of mouse, and again reaches a plateau where the GC distribution remains unchanged (cf. Fig. 1). Also shown are the entire sequence of chromosome 21 in human, and two halves of a long, GC-poor isochore that extends over ~7 Mb of this chromosome. The strong compositional homogeneity, and near identity, of the two halves is striking, compared to the much higher heterogeneity of the entire chromosome (at the time of writing, the mouse draft sequence is still lacking one of the long chromosomes, so it is not shown).

In Fig. 2 it can also be seen that the mouse plot is substantially lower than the human plot. In other words, although the GC distributions of human and mouse share some similarities, the mouse distribution is distinctly narrower than the human one, at equal molecular weights. The profile difference between human and mouse represents the best-documented compositional deviation among eutherian (placental) mammals, which affected some, but not all, rodents. The narrower profiles of the affected rodents, which include murids (e.g., mouse, rat and hamster), differ markedly from those of human, cow, guinea pig and many other eutherians: the GC-richest and GC-poorest DNA appears to have been "eroded", or rather shifted inwards toward the mean of the GC distribution. As a result of the erosion on the steep GC-poor side, the mode is found closer to the mean, i.e. the modal GC is higher and the asymmetry is lower than in other eutherians. A recent CsCl analysis (Douady et al. 2000) has improved the taxonomic resolution of the rodents

that were affected by this compositional inward shift (see below).

Using CsCl profiles to study compositional differences among vertebrate orders

The narrower absorbance profiles of mouse reflect a less pronounced contrast in GC between the GC-richest and GC-poorest isochores in its genome (compared, for example, to human; an illustrative comparison for the major histocompatibility locus is given in Pavlíček et al. 2002b). The narrower profiles also reflect a narrowing in the GC_3 distribution of mouse genes. Such changes are related to the higher substitution rate (Li 1997) and less meticulous DNA repair and maintenance (Holliday 1995) that have been documented in mouse. Other, functionally relevant contrasts within the mammalian genome, such as the contrast between unmethylated CpG islands and other regions of the genome, are similarly reduced in mouse (Douady et al. 2000 and refs. therein).

The inward shift found in some rodents has been called the minor shift (reviewed in Bernardi 2000b), to distinguish it from the much larger, major shift that occurred during the evolution leading to warm-blooded vertebrates, and that distinguishes mammals and birds from most extant cold-blooded vertebrates. The situation is illustrated by some recently obtained CsCl/AUC data, shown in Fig. 3. The clearest difference in this figure (birds are not shown) is between eutherian mammals and fishes. Whereas eutherians (apart from certain rodents) have conserved a high heterogeneity, fishes cluster around much lower values (cf. Bucciarelli et al. 2002). While many reptiles, and especially fishes and amphibians, have narrower CsCl profiles than most mammals, the widths of snake CsCl profiles span a wide range (Hughes et al. 2002).

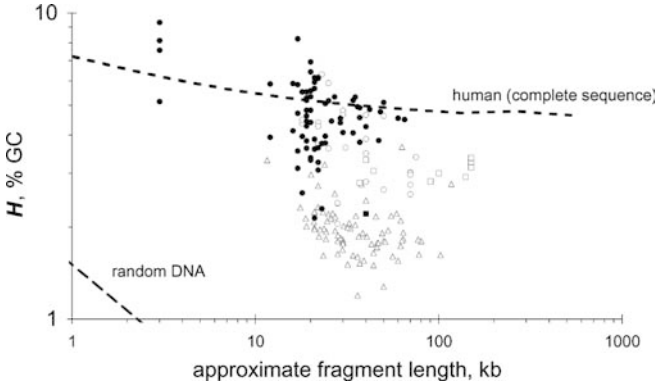

Fig. 3 Compositional heterogeneity H of vertebrate DNAs, as estimated directly from absorbance profiles that were obtained by analytical ultracentrifugation. The CsCl profiles were fitted automatically using a truncated exponential distribution with cubic tail correction; only good fits are included. H^2 was then calculated (see Theory) from the fit's total variance σ^2 by subtracting an estimate of the diffusion contribution, (44.5 kb)/l. Since there is no known procedure for normalizing GC heterogeneity of entire genomes with respect to fragment length l, this length is shown as a second variable, as in Fig. 2. Data are from the following sources: mammals (*closed circles*): C.J. Douady et al. (manuscript in preparation) and Douady et al. (2000); snakes (*open circles*), other reptiles (*open squares*) and amphibian (*closed square*): Hughes et al. (2002); fishes (*triangles*): Bucciarelli et al. (2002). Random and human sequence lines are as in Fig. 2. The lowest mammalian heterogeneities in the plot are for marsupials (noneutherians) and for some rodents. Birds (not shown) have high heterogeneities that can exceed those of most mammals (Thiery et al. 1976; Kadi et al. 1993)

Using CsCl profiles to study compositional differences within a mammalian order

In a study to examine the extent of the minor shift in rodents, we compared a taxonomically diverse set of rodents (Douady et al. 2000). Because of the nature of the minor shift, which corresponds to a sizable inward shift (or "erosion") of GC-rich as well as of GC-poor DNA, the displaced mode is informative for this shift. It is especially useful together with a second parameter that indicates the shape or width of the profile. Figure 4 shows a scatterplot for profile mode and asymmetry of the rodent species we studied, expressed (as above) in GC% instead of buoyant density, for simplicity. From this scatterplot one can see the approximate extent of the minor shift, which affects all of the infraorder Myodonta; more sampling will be needed to clarify whether the Gliridae are also affected.

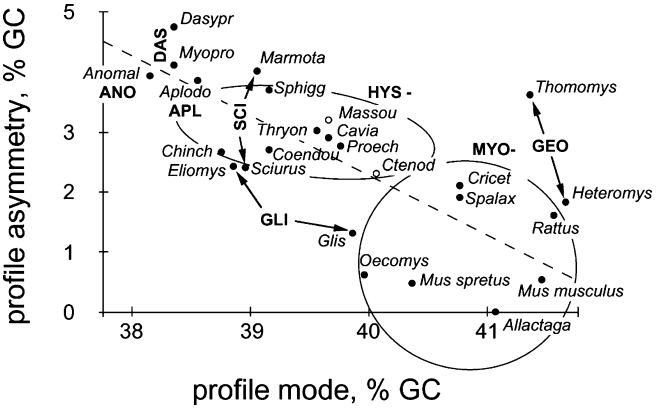

Fig. 4 Asymmetry (mean−mode difference) versus mode scatterplot summarizing rodent CsCl profiles (GC distributions). While almost all non-myomorph rodents have asymmetries of 2% GC or higher, many myomorph rodents (notably murids such as rat, mouse and hamster) can have low asymmetries, as well as low compositional heterogeneities and high modes (see text). Genera are labelled by the first six letters of their Latin names, and main taxonomic groups by three-letter abbreviations: ANO = Anomaluridae, APL = Aplodontidae, DAS = Dasyproctidae, GEO = Geomyoidea, GLI = Gliridae, HYS^- = Hystricognathi minus Dasyproctidae but including Ctenodactylidae gundis (*open circles*), MYO^- = Myodonta (Myomorpha minus Gliridae), SCI = Sciuridae. Modified from Douady et al. (2000), where details are given. Most molecular weights were 50 kb or higher, and all were above 40 kb

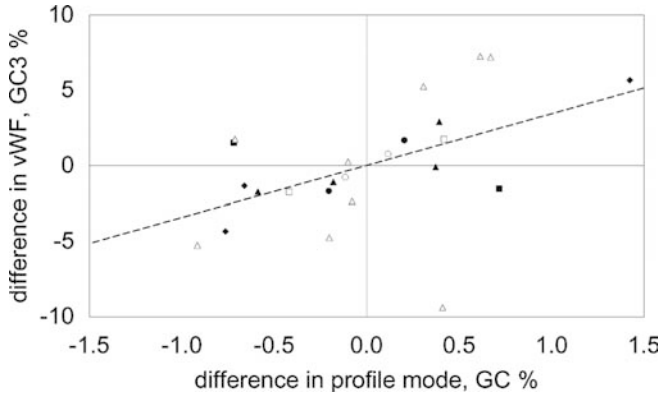

Fig. 5 Intraordinal shifts of CsCl profile modes during mammalian evolution (obtained by AUC) correlate with GC_3 shifts of a 600-bp coding sequence, which is used here as a marker of GC-rich genes. The sequence is from exon 28 of the von Willebrand factor (vWF) gene (Madsen et al. 2001; $GC_3 = 85.5\%$ in human). The correlation coefficient is $R = 0.50$, or 0.59 if the (composite) bat order is excluded. *Points* represent species wherever possible, or else closely related species. Murids were excluded, because of the well-documented and unusual distortion of their profiles (see text). Deviations from the mean value of the order are shown, for the vWF sequence (ordinate) and for the profile mode (abscissa). Orders (Madsen et al. 2001) and species: carnivores (*closed circles*): dog, cat; even-toed ungulates (*open circles*): hippopotamus, camel/llama; bats (*closed squares*): *Dobsonia* megabat, *Tonatia*/*Micronycteris* microbat; primates (*open squares*): human, spider/howler monkey; Afrotheria (*closed triangles*): elephant, aardvark, elephant shrew, hyrax; non-murid rodents (*open triangles*): guinea pig (*outlier*), agouti, New World porcupine, cane rat, mountain beaver, forest/mouse-tailed dormouse, fat dormouse, flying squirrel, woodchuck; lagomorphs (lozenges): rabbit, cottontail, hare

As a cautionary note for this and other considerations, it should be mentioned that buoyant densities of satellite DNAs do not always obey the usual relation to GC level (Corneo et al. 1968): such DNAs, if undetected where they cryptically hide in the main profile, may give incorrect GC estimates and render GC distributions inaccurate. Well-documented abundances of satellite DNA exist within the Geomyoidea superfamily, for example, so that the parameters for its two taxa represented in Fig. 4 (*Thomomys* and *Heteromys*) may be less reliable. It is partly for such reasons that many CsCl studies report only buoyant densities, rather than labelling them by their (usually) equivalent GC levels. For clarity and to facilitate comparison with DNA sequence data, we have shown equivalent GC levels in all figures and equations of this article.

From Fig. 4 we also see that the similarities between the parameters of different taxa correspond well to their grouping according to current criteria (see, e.g. McKenna and Bell 1997, from which the infraorders and superfamilies are taken). This fact indicates also the potential of CsCl analyses for clarifying unresolved phylogenies and confirming phylogenetic hypotheses.

A concordance between the GC levels of individual genes and whole-genome ultracentrifugation results is illustrated in Fig. 5. Such concordances are a result of linear relationships (correlations) that are maintained, in mammalian genomes, between the levels in third codon positions of genes (GC_3) and the GC levels of the much longer DNA regions that embed them. In the human genome, this linear relation was first estimated by hybridizing sequenced genes and measuring the GC levels of the corresponding fragments by ultracentrifugation, and has now been confirmed using the draft sequence: estimates of the relation are consistently close to $GC_3\% \approx 2.92\ GC\% - 74.3$ (Zoubak et al. 1996). An early use of the relation was to deduce the gene density distribution in the human genome (see Introduction, and Fig. 1). More simply, this relation implies that GC_3-rich sequences will typically be found in the GC-rich flank of the CsCl profile. Thus, shifts of the CsCl profile during the evolution of mammalian taxa can sometimes be tracked by sequencing a GC-rich marker gene in different species, and vice versa, as is illustrated here.

Acknowledgements We thank Salvatore Bocchetti for expert work on the XL-A, Giacomo Bernardi, Adam Pavlíček, Gabriel Macaya, Carl W. Schmid, Kamel Jabbari, Francois Catzeflis, Ralph Chamberlin, Laura Giangiacomo, Tom Laue and Peter Schuck for helpful discussions and information, John E. Hearst for clarifications, Stéphane Cruveiller and Borries Demeler for XL-A and software advice, and many people for contributing DNA samples. This work was funded by TMR grant FRMX-CT98-0221 from the European Community.

References

Aparicio S, Chapman J, Stupka E, et al. (2002) Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science 297:1301–1310

Berg HC (1983) Random walks in biology, 2nd edn. Princeton University Press, Princeton, NJ

Bernardi G (2000a) Isochores and the evolutionary genomics of vertebrates. Gene 241:3–17

Bernardi G (2000b) The compositional evolution of vertebrate genomes. Gene 259:31–43

Bernardi G (2001) Misunderstandings about isochores. Part I. Gene 276:3–13

Bernardi G, Bernardi G (1986) Compositional constraints and genome evolution. J Mol Evol 24:1–11

Bernardi G, Bernardi G (1990) Compositional patterns in the nuclear genome of cold-blooded vertebrates. J Mol Evol 31:265–281

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F (1985) The mosaic genome of warm-blooded vertebrates. Science 228:953–958

Bettecken T, Aissani B, Müller C, Bernardi G (1992) Compositional mapping of the human dystrophin-encoding gene. Gene 122:329–335

Bucciarelli G, Bernardi G, Bernardi G (2002) An ultracentrifugation analysis of two hundred fish genomes. Gene 295:153–162

Clay O (2001) Standard deviations and correlations of GC levels in DNA sequences. Gene 276:33–38

Clay O, Carels N, Douady C, Macaya G, Bernardi G (2001) Compositional heterogeneity within and among isochores in mammalian genomes. I. CsCl and sequence analyses. Gene 276:15–24

Corneo G, Ginelli E, Soave C, Bernardi G (1968) Isolation and characterization of mouse and guinea pig satellite DNAs. Biochemistry 7:4373–4379

Cuny G, Soriano P, Macaya G, Bernardi G (1981) The major components of the mouse and human genomes. I. Preparation,

basic properties and compositional heterogeneity. Eur J Biochem 115:227–233
De Sario A, Geigl E-M, Bernardi G (1995) A rapid procedure for the compositional analysis of yeast artificial chromosomes. Nucleic Acids Res 23:4013–4014
De Sario A, Geigl E-M, Palmieri G, D'Urso M, Bernardi G (1996) A compositional map of human chromosome band Xq28. Proc Natl Acad Sci USA 93:1298–1302
Dessauer HC, Coles CJ, Hafner MS (1996) Collection and storage of tissues. In: Hillis DM, Moritz C, Mable BK (eds) Molecular systematics, 2nd edn. Sinauer, Sunderland, Mass., pp 29–47
Douady C, Carels N, Clay O, Catzeflis F, Bernardi G (2000) Diversity and phylogenetic implications of CsCl profiles from rodent DNAs. Mol Phylogenet Evol 17:219–230
Eigner J, Doty P (1965) The native, denatured and renatured states of deoxyribonucleic acid. J Mol Biol 12:549–580
Filipski J, Thiery JP, Bernardi G (1973) An analysis of the bovine genome by $Cs_2SO_4^-Ag^+$ density gradient centrifugation. J Mol Biol 80:177–197
Fujita H (1962) Mathematical theory of sedimentation analysis. Academic Press, New York
Hearst JE, Schmid CW (1973) Density gradient sedimentation equilibrium. Methods Enzymol 27:111–127
Hearst JE, Ifft J, Vinograd J (1961) The effects of pressure on the buoyant behavior of deoxyribonucleic acid and tobacco mosaic virus in a density gradient at equilibrium in the ultracentrifuge. Proc Natl Acad Sci USA 47:1015–1025
Holliday R (1995) Understanding ageing. Cambridge University Press, Cambridge
Hughes S, Clay O, Bernardi G (2002) Compositional patterns in reptilian genomes. Gene 295:323–329
Ifft J, Voet D, Vinograd J (1961) The determination of density distributions and density gradients in binary solutions at equilibrium in the ultracentrifuge. J Phys Chem 65:1138–1145
IHGSC (International Human Genome Sequencing Consortium) (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921
Jabbari K, Cacciò S, Pa de Barros J, Desgrès J, Bernardi G (1997) Evolutionary changes in CpG and methylation levels in the genome of vertebrates. Gene 205:109–118
Kadi F, Mouchiroud D, Sabeur G, Bernardi G (1993) The compositional patterns of the avian genomes and their evolutionary implications. J Mol Evol 37:544–551
Kirk JT (1967) Effect of methylation of cytosine residues on the buoyant density of DNA in caesium chloride solution. J Mol Biol 28:171–172
Li W (1997) Molecular evolution. Sinauer, Sunderland, Mass
Li W (2001) Delineating relative homogeneous G + C domains in DNA sequences. Gene 276:57–72
Macaya G, Thiery JP, Bernardi G (1976) An approach to the organization of eukaryotic genomes at a macromolecular level. J Mol Biol 108:237–254
Madsen O, Scally M, Douady C, Kao D, Debry R, Adkins R, Amrine H, Stanhope M, De JongW, Springer M (2001) Parallel adaptive radiation in two major clades of placental mammals. Nature 409:610–614
Marmur J, Doty P (1959) Heterogeneity in deoxyribonucleic acids: I. Dependence on composition of the configurational stability of deoxyribonucleic acids. Nature 183:1427–1429
McKenna M, Bell S (1997) Classification of mammals above the species level. Columbia University Press, New York
Meselson M, Stahl F, Vinograd J (1957) Equilibrium sedimentation of macromolecules in density gradients. Proc Natl Acad Sci USA 43:581–588
Mouchiroud D, D'Onofrio G, Aissani B, Macaya G, Gautier C, Bernardi G (1991) The distribution of genes in the human genome. Gene 100:181–187
Oliver J, Bernaola-Galván P, Carpena P, Román-Roldán R (2001) Isochore chromosome maps of eukaryotic genomes. Gene 276:47–56
Pavlíček A, Paces J, Clay O, Bernardi G (2002a) A compact view of isochores in the draft human genome sequence. FEBS Lett 511:165–169
Pavlíček A, Clay O, Jabbari K, Paces J, Bernardi G (2002b) Isochore conservation between MHC regions on human chromosome 6 and mouse chromosome 17. FEBS Lett 511:175–177
Rolfe R, Meselson M (1959) The relative homogeneity of microbial DNA. Proc Natl Acad Sci USA 45:1039–1043
Saccone S, Federico C, Bernardi G (2002) Localization of the gene-richest and the gene-poorest isochores in the interphase nuclei of mammals and birds. Gene 300:169–178
Samal S, Geckeler K (2001) Unexpected solute aggregation in water on dilution. Chem Commun 2224–2225
Schildkraut CL, Marmur J, Doty P (1962) Determination of the base composition of deoxyribonucleic acid from its buoyant density in CsCl. J Mol Biol 4:430–443
Schmid CW, Hearst JE (1969) Molecular weights of homogeneous coliphage DNAs from density-gradient sedimentation equilibrium. J Mol Biol 44:143–160
Schmid CW, Hearst JE (1971) Density-gradient sedimentation equilibrium of DNA and the effective density gradient of several salts. Biopolymers 10:1901–1924
Schmid CW, Hearst JE (1972) Sedimentation equilibrium of DNA samples heterogeneous in density. Biopolymers 11:1913–1918
Sueoka N (1959) A statistical analysis of deoxyribonucleic acid distribution in density gradient centrifugation. Proc Natl Acad Sci USA 45:1480–1490
Sueoka N, Marmur J, Doty P (1959) Heterogeneity in deoxyribonucleic acids. II. Dependence of the density of deoxyribonucleic acids on guanine-cytosine content. Nature 183:1429–1433
Szybalski W (1968) Use of cesium sulfate for equilibrium density gradient ultracentrifugation. Methods Enzymol 12:330–360
Tenzen T, Yamagata T, Fukagawa T, Sugaya K, Ando A, Inoko H, Gojobori T, Fujiyama A, Okumura K, Ikemura T (1997) Precise switching of DNA replication timing in the GC content transition area in the human major histocompatibility complex. Mol Cell Biol 17:4043–4050
Thiery JP, Macaya G, Bernardi G (1976) An analysis of eukaryotic genomes by density gradient centrifugation. J Mol Biol 108:219–235
van Holde KE, Johnson WC, Ho PS (1998) Principles of physical biochemistry. Prentice-Hall, Upper Saddle River, NJ
Zoubak S, Clay O, Bernardi G (1996) The gene distribution of the human genome. Gene 174:95–102