Bulletin of Mathematical Biology (2008) 70: 643–653
DOI 10.1007/s11538-007-9261-6

ORIGINAL ARTICLE

Nucleotide Frequencies in Human Genome and Fibonacci Numbers

Michel E. Beleza Yamagishi[a,b,*], Alex Itiro Shimabukuro[c]

[a] *Embrapa Informática, Laboratório de Bioinformática Aplicada, Av. Andre Tosello, 209, Campinas, SP, Brazil*
[b] *Centro Universitário Salesiano de São Paulo—UNISAL, Curso de Ciência da Computacão, Av. Almeida Garret, 267, Campinas, SP, Brazil*
[c] *PUC Campinas—CEATEC, Rodovia D. Pedro I, km 136, Campinas, SP, Brazil*

Received: 14 November 2006 / Accepted: 20 July 2007 / Published online: 10 November 2007

Abstract This work presents a mathematical model that establishes an interesting connection between nucleotide frequencies in human single-stranded DNA and the famous Fibonacci's numbers. The model relies on two assumptions. First, Chargaff's second parity rule should be valid, and second, the nucleotide frequencies should approach limit values when the number of bases is sufficiently large. Under these two hypotheses, it is possible to predict the human nucleotide frequencies with accuracy. This result may be used as evidence to the Fibonacci string model that was proposed to the sequence growth of DNA repetitive sequences. It is noteworthy that the predicted values are solutions of an optimization problem, which is commonplace in many of nature's phenomena.

Keywords Chargaff's parity rules · Nucleotide frequencies · Fibonacci numbers · Golden ratio · Repetitive sequences · Optimization problem

1. Introduction

The amount of available genome data is increasing very fast due to the completion of a host of genome sequencing projects. The careful analysis of all these data is only beginning. The genome sequence by itself is meaningless, it is necessary to identify genes, proceed the annotation, and, if possible, get some understanding of the very process responsible by the sequence formation.

Less than 25% of the fly genome is in coding regions, and the number falls to less than 3% in humans (Do and Choi, 2006). It seems that the most part of eukaryotes genomes is "junk" DNA. Nevertheless, recently, some evidence shows that it is not the case. Mutations in noncoding regions were associated with cancer (Schwartz et al., 2006). Consequently, the interest in noncoding regions has increased, and the role that those regions have in the whole genome demands a better comprehension.

*Corresponding author.
E-mail address: michel@cnptia.embrapa.br (Michel E. Beleza Yamagishi).

Some unexpected facts were revealed after the genome sequences analysis. For instance, almost half of the mammalian genome is composed of repetitive elements (Bannert and Kurth, 2004). Repetitive elements, or simply repeats, can be subdivided into those that are tandemly arrayed and those that are interspersed. Examples of the first class are microsatellites, minisatellites and telomeres; examples of the second class are the transposable elements or transposons. Such considerable amount of repeats should play some role in the living organisms. Actually, the influence of transposons present in the human germ line on gene expression can be envisages by the fact that roughly one quarter of all analyzed human promoters regions harbor sequences derived from these elements (Jordan et al., 2003). However, the most part of the repetitive elements lacks any recognizable function, and seems to be part of the famous "junk" DNA. It is interesting to notice that in terms of those eukaryotes that have their genomes completed, there is an approximately linear correlation between genome size and the total number of DNA repetitive elements present, although the contribution is more significant in larger genomes (Gregory, 2005).

It is possible to propose mathematical models to the tandem repeats sequence growth. Possibly the simplest is based on Fibonacci strings (Dress et al., 2003). In this model, the words are considered over the genetic alphabet and start with two arbitrary strings. These strings or fragments of them may be excised and become candidates for replication and concatenation, and this growth process can go on iteratively by the recurrence formula:

$$\begin{aligned} &S_{a,b}(0) = a; \qquad S_{a,b}(1) = b; \\ &S_{a,b}(k+1) = S_{a,b}(k) + S_{a,b}(k-1) \quad (k \geq 1), \end{aligned} \tag{1}$$

where a and b are arbitrary start strings, and the plus signal, $+$, is supposed to mean string concatenation.

Of course, some error may be introduced in the concatenation step what turns the problem of recognition of this process really difficult. However, if this process really takes place, then some kind of signature should be left behind.

The initial step in any genome analysis is to perform some simple statistical measures like frequencies and averages. This kind of research has been done even before the discovery of DNA structure, and allowed some striking scientific advances. For instance, in Chargaff (1951) observed that, in any piece of double-stranded DNA, the frequencies of adenine and thymine are equal, and so are the frequencies of cytosine and guanine. In mathematical notation $P_A = P_T$ and $P_C = P_G$, where P_A, P_C, P_G and P_T denote the nucleotide frequencies of adenine, cytosine, guanine and thymine, respectively. This observation is known as *Chargaff's first parity rule*. Watson and Crick (1953), were acquainted with Chargaff's first parity rule, and used it to support their DNA double-helix model. Furthermore, Chargaff also observed that the parity rule approximately holds in a single-stranded DNA, nonetheless, the equality is not strict, but $P_A \cong P_T$ and $P_C \cong P_G$. This is known as *Chargaff's second parity rule*. Possibly, the best explanation to this rule can be found in Forsdyke and Bell (2004). Chargaff's second rule has been extensively tested (Mitchell and Bridge, 2006) and proved to hold in the majority of the genome sequences.

A particular interesting case is found in human genome. We have tested the Chargaff's second parity rule for each one of the 24 human chromosomes ($22 + X + Y$), and it is definitely valid. Moreover, it is known that $P_A + P_T + P_C + P_G = 1$ by definition (the sum of all frequencies must be equal to 1), and assuming *Chargaff's second parity rule*,

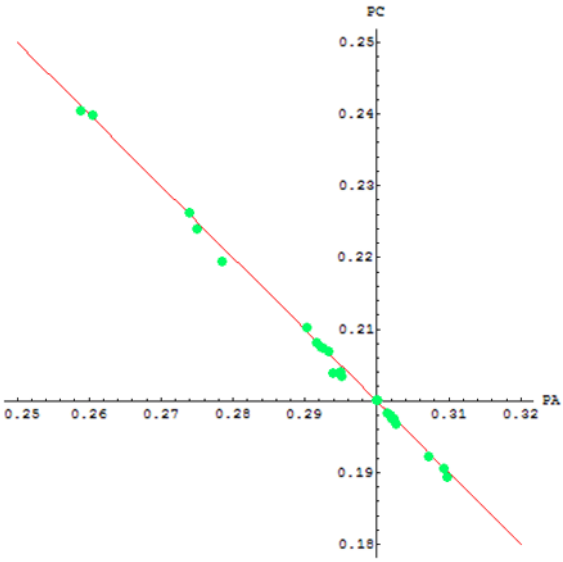

Fig. 1 In *red*, the *line* $P_A + P_C = \frac{1}{2}$, and in *green*, the observed points (P_A, P_C) for each human chromosome (colour figure online).

we get that $P_A + P_C \cong \frac{1}{2}$ or, equivalently, $P_T + P_G \cong \frac{1}{2}$, or any possible combination. If we plot the points (P_A, P_C) for each human chromosome, we get another interesting fact: they are not evenly spread over the line $P_A + P_C = \frac{1}{2}$, but seem to be aggregated around some very precise values. In Fig. 1, in red, the line $P_A + P_C = \frac{1}{2}$, and the green dots are the points (P_A, P_C) for each human chromosome.

Although this observation is not expected, it is not completely unusual. Many phenomena in nature show the same pattern, and some of them can be mathematically modeled. Usually, those mathematical models that describe nature's phenomena involve optimization problems. It seems that nature is always trying to optimize itself in different contexts. Therefore, the following question emerges naturally: Is it possible to build a mathematical model that predicts or explain the observed frequencies?

Assuming that (i) the human nucleotide frequencies really tend to limit values when the number of bases is sufficiently large, and (ii) Chargaff's second parity rule is valid, we derived a mathematical model that predicts the observed frequency values with accuracy.

2. Mathematical model

In order to understand our model, it is necessary to introduce the Fibonacci numbers (Fibonacci and Singler, 2002).

2.1. Fibonacci numbers

In mathematics, one of the most famous integer sequence is without doubt the sequence $\{1, 1, 2, 3, 5, 8, 13, \ldots\}$.

This sequence, called Fibonacci sequence, is obtained through the recurrence formula

$$F(n+2) = F(n+1) + F(n), \tag{2}$$

together with the initial conditions $F(1) = 1$ and $F(2) = 1$.

The Fibonacci sequence was first described, in the Occident, by *Leonardo of Pisa*, also known as Fibonacci, in his book *Liber Abaci*. The Fibonacci sequence appears in nature in different contexts: sea shell shapes, flower petals and seeds, etc.

It is related to the *Golden Ratio*, ϕ, by the limit

$$\phi = \lim_{n\to\infty} \frac{F(n+1)}{F(n)}. \tag{3}$$

The Golden Ratio is associated to Beauty and Perfection, and for this reason, it is conventional to find ϕ present in art (Leonardo da Vinci), architecture (Parthenon in Athens, for example) and music (notably in Bartók and Debussy). There is a plenty of written works about the Golden Ratio and the Fibonacci numbers.

2.2. Assumptions and model

The main assumption of our model is that Chargaff's second parity rule is valid in all human chromosomes. There are many different forms to state it mathematically. We've decided to do it in the following way: *the division of the frequency of one nucleotide by the sum of the frequencies of the remaining nucleotides is in the proportion of three Fibonacci numbers.* The choice of Fibonacci numbers were based not only on their generalized occurrence in nature, but also taking in account that the growth process of repetitive sequences may be related to Fibonacci strings in some sense. It is also noteworthy that we've also tried other sets of numbers, but with no success.

Consider the three Fibonacci numbers below

$$\{F(n), F(n+1), F(n+k)\}, \tag{4}$$

where n is a sufficiently large number and $k = 0, 1, 2, 3, \ldots, N$ (N is finite number).

Therefore, we can write the main assumption as

$$\frac{x(n)}{y(n)+z(n)+w(n)} \propto \frac{F(n)}{F(n+k)}, \tag{5}$$

$$\frac{y(n)}{x(n)+z(n)+w(n)} \propto \frac{F(n+1)}{F(n+k)}, \tag{6}$$

$$\frac{z(n)}{x(n)+y(n)+w(n)} \propto \frac{F(n)}{F(n+k)}, \tag{7}$$

$$\frac{w(n)}{x(n)+y(n)+z(n)} \propto \frac{F(n+1)}{F(n+k)}, \tag{8}$$

where $x(n)$, $y(n)$, $z(n)$ and $w(n)$ represent the nucleotide frequencies, without any a priori association, when the number of nucleotide bases is n, i.e., $x(n) = \frac{x_n}{n}$, where x_n stands for the number of nucleotide x.

It is not straightforward to recognize Chargaff's second parity rule in Eqs. (5)–(8). One way to grasp the idea behind the formulas is to note that Eqs. (5) and (7) are proportional to the same quotient ($\frac{F(n)}{F(n+k)}$), and the same can be said about Eqs. (6) and (8). In next section, we will show how to get Chargaff's second parity rule from the above equations.

2.2.1. Limit values

Now, lets impose our second assumption, i.e., that the nucleotide frequencies tend to limit values when n is sufficiently large. Mathematically, it can be written as

$$x = \lim_{n\to\infty} \frac{x_n}{n}, \tag{9}$$

$$y = \lim_{n\to\infty} \frac{y_n}{n}, \tag{10}$$

$$z = \lim_{n\to\infty} \frac{z_n}{n}, \tag{11}$$

$$w = \lim_{n\to\infty} \frac{w_n}{n}. \tag{12}$$

It is also necessary to understand what happens with the quotients $\frac{F(n)}{F(n+k)}$ and $\frac{F(n+1)}{F(n+k)}$ when $n \to \infty$.

Using Eq. (2) recursively, it is easy to get the following recurrence formula

$$F(n+k) = F(k)F(n+1) + F(k-1)F(n). \tag{13}$$

We are particularly interested in the cases where n, the numbers of bases, is large, and the quotient of the Fibonacci numbers tends to a limit.

Mathematically, this can be obtained as follows. Dividing (13) by $F(n+k)$, we get

$$1 = F(k)\frac{F(n+1)}{F(n+k)} + F(k-1)\frac{F(n)}{F(n+k)}. \tag{14}$$

Taking the limit as $n \to \infty$,

$$1 = F(k)\lim_{n\to\infty}\frac{F(n+1)}{F(n+k)} + F(k-1)\lim_{n\to\infty}\frac{F(n)}{F(n+k)}. \tag{15}$$

We define

$$\lambda_{1,k} = \lim_{n\to\infty}\frac{F(n+1)}{F(n+k)} \tag{16}$$

and

$$\lambda_{2,k} = \lim_{n\to\infty}\frac{F(n)}{F(n+k)}. \tag{17}$$

Observe that $\lambda_{1,k}$ and $\lambda_{2,k}$ are linked to the Golden Ratio by

$$\lambda_{1,k} = \phi^{1-k} \tag{18}$$

and

$$\lambda_{2,k} = \phi^{-k}, \tag{19}$$

respectively. It is interesting that in Dress et al. (2003) the exponential growth of the sequence length of repetitive sequences in the Fibonacci–Cayley model is the order of ϕ^k, where k is related to the kth step in the Fibonacci strings concatenation process.

Thus, the Eq. (15) can be written as

$$1 = F(k)\lambda_{1,k} + F(k-1)\lambda_{2,k}. \tag{20}$$

Finally, our model can be rewritten as

$$\frac{x}{y+z+w} = \lambda_{1,k}, \tag{21}$$

$$\frac{y}{x+z+w} = \lambda_{2,k}, \tag{22}$$

$$\frac{z}{x+y+w} = \lambda_{1,k}, \tag{23}$$

$$\frac{w}{x+y+z} = \lambda_{2,k}. \tag{24}$$

As noted before, x, y, z, and w are frequencies, so we have

$$x + y + z + w = 1. \tag{25}$$

Using Eq. (25), the Eqs. (21)–(24), we get

$$\frac{x}{1-x} = \lambda_{1,k}, \tag{26}$$

$$\frac{y}{1-y} = \lambda_{2,k}, \tag{27}$$

$$\frac{z}{1-z} = \lambda_{1,k}, \tag{28}$$

$$\frac{w}{1-w} = \lambda_{2,k}. \tag{29}$$

Equations (26) and (28) imply that

$$x = z \tag{30}$$

and, analogously, Eqs. (27) and (29) imply that

$$y = w. \tag{31}$$

Equations (30) and (31) are the Chargaff's second parity rule.

Moreover, an immediate consequence of Eqs. (30), (31) and (25) is

$$x + y = \frac{1}{2}. \tag{32}$$

2.3. *Optimization problem*

Now, we have 3 equations in two variables

$$\frac{x}{1-x} = \lambda_{1,k}, \tag{33}$$

$$\frac{y}{1-y} = \lambda_{2,k}, \tag{34}$$

$$x + y = \frac{1}{2}, \tag{35}$$

which can be rewritten as

$$x = \frac{\lambda_{1,k}}{1+\lambda_{1,k}}, \tag{36}$$

$$y = \frac{\lambda_{2,k}}{1+\lambda_{2,k}}, \tag{37}$$

$$x + y = \frac{1}{2}. \tag{38}$$

This is a linear system, and, using Eqs. (20) and (32), it is not difficult to show that it is inconsistent, independently, of k. In fact, only when $k \to \infty$, the system is consistent, but we are dealing with the cases where k is finite.

The Eq. (32) must be satisfied because x and y are frequencies and, by definition, the Eq. (25) must hold. Therefore, we should try to minimize the difference between x and $\frac{\lambda_{1,k}}{1+\lambda_{1,k}}$, and the difference between y and $\frac{\lambda_{2,k}}{1+\lambda_{2,k}}$ under the condition that $x + y = \frac{1}{2}$.

This is a classical optimization problem, and can be mathematically stated as

$$\min_{x+y=\frac{1}{2}} f_k(x, y), \tag{39}$$

where

$$f_k(x, y) = \left(x - \frac{\lambda_{1,k}}{1+\lambda_{1,k}}\right)^2 + \left(y - \frac{\lambda_{2,k}}{1+\lambda_{2,k}}\right)^2. \tag{40}$$

This minimization problem is sufficiently easy to solve, because its objective function is quadratic and the Jacobian of the constraint is full rank, therefore, the solution exists and is unique (Nocedal and Wright, 2000).

In Table 1, we list the solutions to the first 8 values of k. It is not difficult to show that $(x, y) \to (0.25, 0.25)$ as $k \to \infty$.

The values of Table 1 are in agreement with the observed frequencies in human chromosomes. In the next section, we will present the data that supports this mathematical model.

Table 1 Solutions of the optimization problem for different values of k

k	x	$x \cong$	y	$y \cong$
0	$\frac{3+\sqrt{5}}{8+4\sqrt{5}}$	0.3090	$\frac{1+\sqrt{5}}{8+4\sqrt{5}}$	0.1909
1	$\frac{3+\sqrt{5}}{8+4\sqrt{5}}$	0.3090	$\frac{1+\sqrt{5}}{8+4\sqrt{5}}$	0.1909
2	$\frac{127+57\sqrt{5}}{420+188\sqrt{5}}$	0.3027	$\frac{83+37\sqrt{5}}{420+188\sqrt{5}}$	0.1972
3	$\frac{161+72\sqrt{5}}{550+246\sqrt{5}}$	0.2927	$\frac{114+51\sqrt{5}}{550+246\sqrt{5}}$	0.2072
4	$\frac{881+392\sqrt{5}}{3126+1398\sqrt{5}}$	0.2818	$\frac{682+305\sqrt{5}}{3126+1398\sqrt{5}}$	0.2181
5	$\frac{20583+9205\sqrt{5}}{75588+33804\sqrt{5}}$	0.2723	$\frac{17211+7697\sqrt{5}}{75588+33804\sqrt{5}}$	0.2276
6	$\frac{15908+7070\sqrt{5}}{59665+26683\sqrt{5}}$	0.2649	$\frac{3(9349+4181\sqrt{5})}{119330+53366\sqrt{5}}$	0.2350
7	$\frac{100793+45076\sqrt{5}}{388045+173539\sqrt{5}}$	0.2597	$\frac{186459+83387\sqrt{5}}{776090+347078\sqrt{5}}$	0.2402

3. Results

We've performed a simple experiment using the nucleotide frequencies in human genome. The human genome data were obtained at NCBI.[1] Every so often, new human genome versions are released. We've used Build 35.1. Two points are worthy to mention. First, only partial data is available for each chromosome, i.e., there are still missing sections, for instance, chromosome 1 is supposed to have about 263 million bases, but only about 220 million bases were available. Second, in a recent study (Salzberg and Yorke, 2005), the very quality of the published genomes was put under doubt. The authors explicitly state that "*Finishing* (*genome*) *efforts are usually directed at closing gaps, not at fixing mis-assemblies, and therefore, 'finished' genomes are very likely to contain errors of the type we are discussing*". The errors discussed in this study were genome mis-assemblies due repetitive sequences. In some cases, the repetitive sequence length is underestimated what really affects our results. It is interesting to notice that the majority of biologists are not aware of this. These two points are relevant because they may explain some minor deviations from the predicted values.

3.1. Human nucleotide frequencies

This experiment consisted in calculating the nucleotide frequencies in all 24 human chromosomes. The results are summarized in Table 2.

The nucleotide frequencies are clustered around the predicted values of Table 1. In Fig. 2, we have in red, the solutions of the optimization problem for different values of k, and in green, the points (P_A, P_C) for each one of the human chromosomes. The red circles have their centers in the solutions of the optimization problem and they have the same radius equal to 0.005.

[1] National Center for Biotechnology Information. Site (http://www.ncbi.nlm.nih.gov).

Table 2 Nucleotide Frequencies for all human chromosomes

Chromosome	P_A	P_C	P_G	P_T	k
Chrom 1	0.2916	0.2080	0.2080	0.2922	3
Chrom 2	0.3000	0.2003	0.2005	0.2997	2
Chrom 3	0.3019	0.1980	0.1980	0.3020	2
Chrom 4	0.3093	0.1905	0.1906	0.3094	1
Chrom 5	0.3020	0.1974	0.1975	0.3011	2
Chrom 6	0.3024	0.1975	0.1976	0.3023	2
Chrom 7	0.2950	0.2040	0.2040	0.2951	3
Chrom 8	0.3002	0.2001	0.2000	0.2999	2
Chrom 9	0.2933	0.2067	0.2067	0.2931	3
Chrom 10	0.2922	0.2074	0.2074	0.2928	3
Chrom 11	0.2925	0.2072	0.2075	0.2926	3
Chrom 12	0.2950	0.2040	0.2033	0.2956	3
Chrom 13	0.3072	0.1922	0.1922	0.3080	1
Chrom 14	0.2951	0.2034	0.2039	0.2974	3
Chrom 15	0.2903	0.2101	0.2099	0.2895	3
Chrom 16	0.2750	0.2040	0.2040	0.2750	4
Chrom 17	0.2740	0.2261	0.2258	0.2713	5
Chrom 18	0.3014	0.1982	0.1985	0.3017	2
Chrom 19	0.2588	0.2403	0.2409	0.2598	7
Chrom 20	0.2785	0.2194	0.2202	0.2817	5
Chrom 21	0.2940	0.2040	0.2039	0.2952	3
Chrom 22	0.2605	0.2398	0.2397	0.2598	7
Chrom X	0.3027	0.1968	0.1967	0.3033	2
Chrom Y	0.3098	0.1893	0.1889	0.3118	1

It is interesting to note that all the points (P_A, P_C) are near (less than 0.005) to the predicted values.

The average values are $\mu_{P_A} = 0.292$, $\mu_{P_C} = 0.207$, $\mu_{P_G} = 0.207$ and $\mu_{P_T} = 0.292$, which are close to the optimization's solution when $k = 3$.

4. Conclusion

Using Chargaff's second parity rule and assuming that the nucleotide frequencies tend to limit values when the number of nucleotide bases is sufficiently large, we've described a mathematical model that predicts the limit values of the human nucleotide frequencies with great accuracy. Our mathematical model maybe be viewed as an evidence that some kind Fibonacci string process might be involved in the DNA sequence growth, particularly, in those DNA repetitive sequences which are almost 50% of the human genome.

It is also interesting to notice that the limit values are the results of an optimization problem, and it is commonly found in many phenomena in nature.

If our two hypotheses hold and our mathematical model is correct, then it is possible to make the following conjecture: the noncoding DNA regions play a major role in the "optimization process" to reach the limit values predicted in our mathematical model. This conjecture is based on the fact that about 97% of human genome is believed to be noncoding.

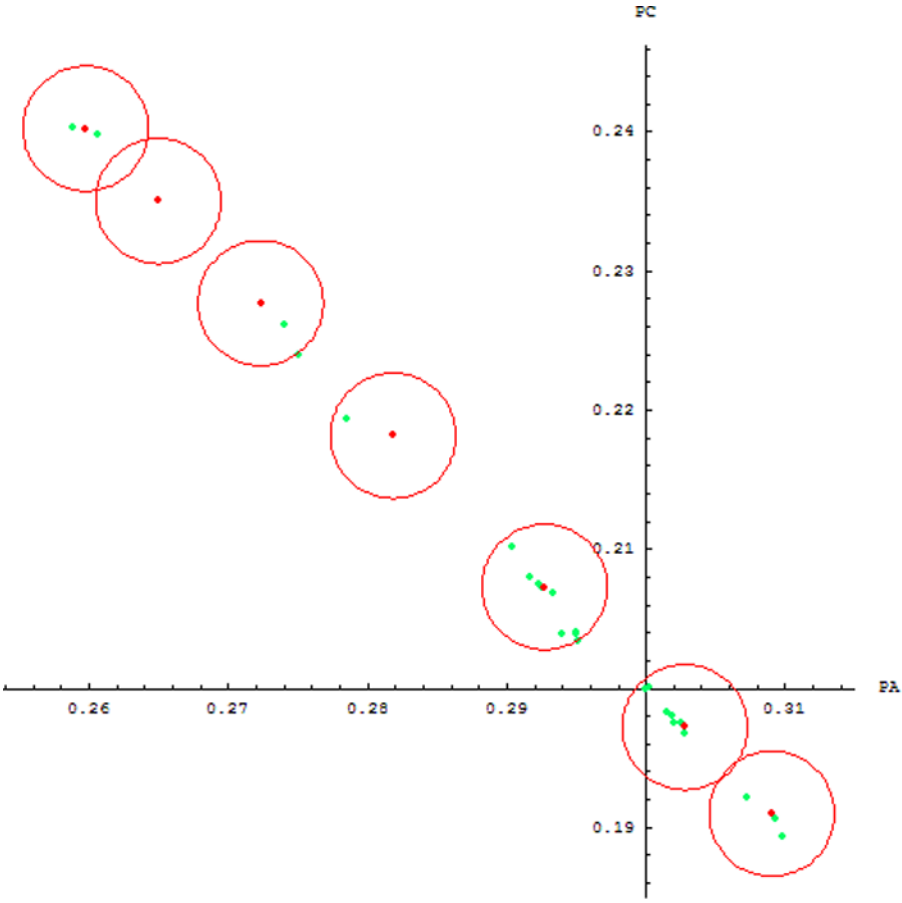

Fig. 2 In *red dots*, the solutions of the optimization problem for different values of k. In *green*, the *points* (P_A, P_C) for each one of the human chromosomes (colour figure online).

Acknowledgements

We are grateful to Dr. Bernard Maigret from Henry Poincaré University (Nancy, France) and Dr. Nir Cohen from Campinas State University (Campinas, Brazil) for their valuable comments on our manuscript. We are also grateful to Dr. Robert Giegerich from Bielefeld University (Bielefeld, Germany) for insightful discussions about our mathematical model. We are also indebted to the anonymous referee for his insightful remarks concerning Fibonacci strings and DNA repetitive sequences.

References

Bannert, N., Kurth, R., 2004. Retroelements and the human genome: new perspectives on an old relation. Proc. Natl. Acad. Sci. 101, 14572–14579.

Chargaff, E., 1951. Structure and function of nucleic acids as cell constituents. Fed. Proc. 10, 654–659.
Dress, A., Giegerich, R., Grunewald, S., Wagner, H., 2003. Fibonacci-Cayley numbers and repetition patterns in genomic DNA. Ann. Comb. 7, 259–279.
Do, J.H., Choi, D.-K., 2006. Computational approaches to gene prediction. J. Microbiol. 44, 137–144.
Fibonacci, L., Singler, L.E., 2002. Fibonacci's Liber Abaci. Springer, New York.
Forsdyke, D.R., Bell, S.J., 2004. A discussion of the application of elementary principles to early chemical observations. Appl. Bioinform. 3, 3–8.
Gregory, T.R., 2005. The C-value enigma in plants and animals: a review of parallels and an appeal for partnership. Ann. Bot. 95, 133–146.
Jordan, I.K., Rogozin, I.B., Glazko, G.V., Koonin, E.V., 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet. 19, 68–72.
Mitchell, D., Bridge, R., 2006. A test of Chargaff's second rule. BBRC 340, 90–94.
Nocedal, J., Wright, S.J., 2000. Numerical Optimization. Springer Series in Operations Research. New York, Springer.
Salzberg, S.L., Yorke, J.A., 2005. Beware of mis-assembled genomes. Bioinformatics 21, 4320–4321.
Schwartz, S., Alazzouzi, H., Perucho, M., 2006. Mutational dynamics in human tumors confirm the neutral intrinsic instability of the mitochondrial D-loop poly-cytidine repeat. Genes Chromosom. Cancer 8, 770–780.
Watson, J.D., Crick, F.H.C., 1953. Molecular structure of nucleic acids. Nature 4356, 737.