BioSystems xxx (2011) xxx–xxx

Contents lists available at SciVerse ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems

Random models of Menzerath–Altmann law in genomes

Jaume Baixeries[a], Antoni Hernández-Fernández[b,c], Ramon Ferrer-i-Cancho[c,*]

[a] Complexity and Quantitative Linguistics Lab, Departament de Llenguatges i Sistemes Informàtics, LARCA Research Group, Universitat Politècnica de Catalunya, Campus Nord, Edifici Omega, Jordi Girona Salgado 1-3, 08034 Barcelona (Catalonia), Spain
[b] Departament de Lingüística General, Universitat de Barcelona, Gran Via de les Corts Catalanes 585, 08007 Barcelona (Catalonia), Spain
[c] Complexity and Quantitative Linguistics Lab, Departament de Llenguatges i Sistemes Informàtics, TALP Research Center, Universitat Politècnica de Catalunya, Campus Nord, Edifici Omega, Jordi Girona Salgado 1-3, 08034 Barcelona (Catalonia), Spain

ARTICLE INFO

Article history:
Received 13 October 2011
Received in revised form
16 November 2011
Accepted 28 November 2011

Keywords:
Random breakage
Power law
Genome size
Chromosome number
Menzerath–Altmann law

ABSTRACT

Recently, a random breakage model has been proposed to explain the negative correlation between mean chromosome length and chromosome number that is found in many groups of species and is consistent with Menzerath–Altmann law, a statistical law that defines the dependency between the mean size of the whole and the number of parts in quantitative linguistics. Here, the central assumption of the model, namely that genome size is independent from chromosome number is reviewed. This assumption is shown to be unrealistic from the perspective of chromosome structure and the statistical analysis of real genomes. A general class of random models, including that random breakage model, is analyzed. For any model within this class, a power law with an exponent of −1 is predicted for the expectation of the mean chromosome size as a function of chromosome length, a functional dependency that is not supported by real genomes. The random breakage and variants keeping genome size and chromosome number independent raise no serious objection to the relevance of correlations consistent with Menzerath–Altmann law across taxonomic groups and the possibility of a connection between human language and genomes through that law.

1. Introduction

Human language, music and genomes share a common statistical pattern: the mean size of the parts tends to decrease as the number of parts increases (Ferrer-i-Cancho and Forns, 2009). On the one hand, it is known in quantitative linguistics that the size of a construct (e.g., a sentence) tends to decrease as the number of constituents (e.g., clauses) increases (Altmann, 1980; Teupenhayn and Altmann, 1984). This behavior has been studied at many linguistic levels: e.g., morphemes, words, sentences (Altmann, 1980; Teupenhayn and Altmann, 1984) and also in music (Boroda and Altmann, 1991). This recurrent statistical pattern is presently known as Menzerath–Altmann law, the mathematical function that is used to define the relationship between the size of a construct and the size of its units (Cramer, 2005). If the mean size of the parts is S_P (e.g., the length of a clause in words), with the size of the whole S_W (e.g., the length of a sentence in clauses), the law states that.

$$S_P \sim S_W{}^b e^{cS_W} \quad (1)$$

where b and c are constants. On the other hand, the genomes of many groups of species exhibit a parallel qualitative behavior: the mean length of chromosomes tends to increase as the number of chromosomes of a species increases (Ferrer-i-Cancho and Forns, 2009; Wilde and Schwibbe, 1989). The significance of this statistical coincidence is a matter of current debate (Hernández-Fernández et al., 2011; Solé, 2010). One of the main arguments that has been raised against the relevance of the finding of patterning consistent with the law in genomes is a simple model of random breakage that can account for a negative correlation between mean chromosome length and chromosome number (Solé, 2010). Here the actual capacity of this model for explaining the real dependency between mean chromosome length and chromosome number will be analyzed. It will be argued that this random breakage model and all the variants within the same class of models are equivalent to a particular case of Menzerath–Altmann law, namely Eq. (1) with $b = -1$ and $c = 0$ and it will be shown that real genomes deviate significantly from Menzerath–Altmann law with precisely these concrete parameters.

* Corresponding author. Tel.: +34 934137870.
E-mail addresses: jbaixer@lsi.upc.edu (J. Baixeries), antonio.hernandez@upc.edu (A. Hernández-Fernández), rferrericancho@lsi.upc.edu (R. Ferrer-i-Cancho).

0303-2647/$ – see front matter © 2011 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2011.11.010

Please cite this article in press as: Baixeries, J., et al., Random models of Menzerath–Altmann law in genomes. BioSystems (2011), doi:10.1016/j.biosystems.2011.11.010

Let us define G as the size of a genome in base pairs, L_g as its number of chromosomes and $L_c = G/L_g$ as its mean chromosome length. It has been argued that Menzerath–Altmann law in genomes could be simply explained through a simple model of random breakage that generates the information about an organism in two stages. In the first stage, the values of G and L_g are generated for a group of N organisms through the following procedure:

- G is chosen uniformly at random within the interval (G^m, G^M).
- L_g is chosen uniformly at random within the interval (L_g^m, L_g^M) and independently from G.
- Given G and L_g, L_c is computed through $L_c = G/L_g$.

In the second stage, the length of each of the L_g chromosomes is generated applying a random breakage procedure (De et al., 2001), i.e. by selecting L_g breaking points along the sequence of length G uniformly at random (Solé, 2010). This procedure is also known as random fragmentation (Sankoff and Ferretti, 1996).

Notice that the independence between G and L_g assumed in the first stage is not a decision that is taken for simplicity but a property that it is claimed to actually hold in real genomes (Solé, 2010). Here it will be argued the opposite, namely that such independence is unrealistic from the perspective of chromosome structure and the statistical properties of actual genomes. It will be shown that the assumption of independence between G and L_g has negative consequences for the suitability of the model for actual genomes. Firstly, the assumption is no supported by real data (Hernández-Fernández et al., 2010) and, under certain parameter choices, can lead to organisms that are not well-formed or viable. Secondly, it will be shown that, due to such independence, $E[L_c|L_g]$, the expectation of L_c given L_g follows a power law with a -1 exponent, i.e.

$$E[Lc|Lg] \sim L_g^{\beta} \tag{2}$$

with $\beta = -1$. This particular power-law (a particular case of Menzerath–Altmann law, i.e. Eq. (1) with $S_P = E[L_c|L_g]$, $S_W = L_g$, $b = -1$ and $c = 0$) will be shown to be insufficient to explain the actual dependency between L_c and L_g.

The remainder of the article is organized as follows. Section 2 evaluates the suitability of the assumption of independence between G and L_g of the random breakage model to real genomes from different perspectives. Section 2.1 argues that there is a structural dependency between G and L_g. A well-formed chromosome must have centromere and two telomeres and this imposes a lower bound on the minimum value of G given L_g. However, this is not a decisive argument against the random breakage model because chromosomes may be so large that this theoretical lower bound may have no observable effect. Section 2.2 reviews the statistical evidence of a dependency between G and L_g that the random breakage model denies. This is still not decisive evidence against the suitability of the random breakage model because the null hypothesis of independence between G and L_g cannot be rejected in two of the major groups of organisms reviewed, birds and cartilaginous fishes. A decisive rejection of the random breakage model (indeed a rejection of a wider class of models) for all the major groups of organisms reviewed arrives in Section 2.3. This subsection shows that

- All the major groups of organisms reviewed (including birds and cartilaginous fishes) deviate significantly from Menzerath–Altmann law with these particular parameters using complementary mathematical and statistical arguments. Indeed, Section 2.3 rejects a generalization of Menzerath–Altmann law with $b = -1$ and $c = 0$ extended by adding an additive term.
- Menzerath–Altmann law with $b = -1$ and $c = 0$ is equivalent to independence between G and L_g.

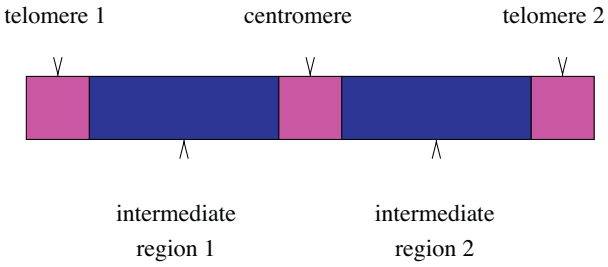

Fig. 1. Scheme of a metacentric chromosome, which has five parts, i.e. two telomeres, one centromere and two intermediate regions between a telomere and the centromere.

- The random breakage model and other models within the same class are equivalent to Menzerath–Altmann law with $b = -1$ and $c = 0$.

Section 2.3 also presents the exact mathematical dependency between L_g and L_c of the random breakage model with uniformly distributed G and L_g (Solé, 2010). Section 3 checks and clarifies the need of independence between G and L_g to obtain Menzerath–Altmann law with $b = -1$ and $c = 0$. This section shows that introducing a structural constraint (i.e. the centromere and the two telomeres must have a minimum size) in genomes that would be formed a priori by choosing chromosome number and genome size independently, leads to a dependency between L_g and L_c that deviates from Menzerath–Altmann law with $b = -1$ and $c = 0$ and whose exact mathematical form is presented. Section 4 summarizes all the problems surrounding Solé's (2010) assumption of independence between genome and chromosome level and explores other pitfalls of his random breakage model.

2. The Problems of Independence Between Genome Size and Chromosome Number

2.1. Well-formed or Viable Chromosomes

The random breakage model above does not impose any constraint on chromosome length but organisms with empty chromosomes are produced when $G < L_g$. In contrast, models of chromosome length evolution (De et al., 2001; Sankoff and Ferretti, 1996) consider that

> "...a viable and functional chromosome must minimally contain a centromere and two telomeres (and at least one gene whose function is not duplicated elsewhere in the genome). This imposes a lower bound on the size of a chromosome, on a purely structural basis. Finally, from the genetic viewpoint, there is reason to believe that for meiosis to be completed successfully, each chromosome must be of length sufficient for at least one crossover to be expected among the four aligned strands before they segregate into two pairs." (Sankoff and Ferretti, 1996, p. 8).

For simplicity, let us assume that a chromosome is made of three main parts: a centromere and two telomeres (Fig. 1). c^m, t_1^m, t_2^m, g_1^m, g_2^m and are defined, respectively as the minimum size of the centromere, the two telomeres and two intermediate regions delimited by a telomere and the centromere. In general, a viable chromosome must satisfy, at least, the condition

$$(c^m + t_1^m + t_2^m + g_1^m + g_2^m)L_g \leq G \tag{3}$$

Following Sankoff and Ferretti (1996), let us consider that a viable genome must contain at least one centromere and two telomeres and that the intermediate regions could be empty. Our course, a further constraint could be added, namely that even the

Please cite this article in press as: Baixeries, J., et al., Random models of Menzerath–Altmann law in genomes. BioSystems (2011), doi:10.1016/j.biosystems.2011.11.010

intermediate regions cannot be empty. Our choice of a general chromosome structure (Fig. 1) is motivated by the goal of showing the structural dependency between the genome and the chromosome scale from a mathematical point of view. Of course, more biologically realistic scenarios can be developed from it.

Assuming c^m, t_1^m, $t_2^m = 1$ and $g_1^m = g_2^m = 0$ as in Sankoff and Ferreti (1997), Eq. (3) yields $G \geq 3L_g$. Applying the parameters of Solé's random breakage model with $g_1^m = g_2^m = 0$, the viability condition in Eq. (3) is always satisfied if

$$(c^m + t_1^m + t_2^m)(L_g^M - 1) = G^m + 1 \quad (4)$$

holds. Next the parameter conditions that warrant that the random breakage model does not produce empty chromosomes or chromosomes with at least one empty part will be derived. Only the two telomeres and the centromere will be considered as main parts.

Empty chromosomes (chromosomes with no base pair) are avoided by imposing $c^m + t_1^m + t_2^m \geq 1$ in Eq. (4), which gives

$$L_g^M \leq G^m + 2, \quad (5)$$

in agreement with Hernández-Fernández et al. (2011). Interestingly imposing the condition that none of the main parts is empty, i.e. c^m, t_1^m, $t_2^m = 1$ on Eq. (4) it is obtained

$$3L_g^M \leq G^m + 4. \quad (6)$$

Notice that the independence between G and L_g needs Eq. (5) if genomes with chromosomes of length zero cannot be generated or Eq. (6) if genomes with empty main parts cannot be produced. Next the results above will be applied to the parameters chosen by Solé (2010) to simulate the random breakage model. It will be shown that in one of the parameter settings, chromosomes with non-viable structure are generated.

The random breakage model was simulated with two parameter settings (Solé, 2010). In both, $N = 500$, $G^M = 10{,}000$, $L_g^m = 10$ and $L_g^M = 200$ were used. In simulation (a) $G^m = 9000$ was used (for Fig. 2(a) of Solé (2010)) and in simulation (b) $G^m = 50$ was used (for Fig. 2(a) of Solé (2010)). While parameter setting (a) satisfies the necessary and sufficient condition for absence of empty chromosomes (Eq. (5)) and the necessary and sufficient condition for absence of chromosomes with empty main parts (Eq. (6)), parameter setting (b) violates both conditions (in this case $L_g^M > G_m + 2$, which implies $3L_g^M > G_m + 4$ as $G^m > 0$). Next we will focus on parameter setting (b), which was used to explain qualitatively the high dispersion in the actual relationship between L_c and L_g in angiosperm plants. The issue of whether $G^m = 50$ (or $G^m = 9000$ in parameter setting (a)) is realistic enough, knowing that a minimum complete living organism has been argued to require about $G^m = 6 \times 10^5$ (Luisi, 2006 and references therein), will be left aside and we will focus on the implications for viability of such as small minimum genome size from a purely structural perspective.

The random breakage model was simulated varying N and using the remainder of parameters from setting (b). Our Fig. 2 shows the growth of the number of organisms with empty chromosomes as N grows. For $N = 500$, 2.9 ± 1.7 organisms with empty chromosomes are produced on average while 13.3 ± 3.6 organisms with at least one chromosome with empty main parts are produced on average. However, if the simulations had taken the value of N of angiosperm plants from the sample studied by Hernández-Fernández et al. (2011), i.e. $N = 4706$, then 27.6 ± 5.2 organisms with empty chromosomes and 125.1 ± 11.0 organisms with at least one empty main chromosome part would be obtained on average.

We have studied the problem of viability by just requiring chromosomes or main chromosome parts to not be empty, which both define necessary but not sufficient conditions for actual chromosome viability as parts of size one may still not be large enough to warrant viability. These simple requirements for chromosome viability have been chosen to illustrate the problem of denying the dependency between G and L_g. A deeper analysis of viability could be performed with more realistic values of c^m, t_1^m, t_2^m, g_1^m, g_2^m or G_m (Luisi, 2006).

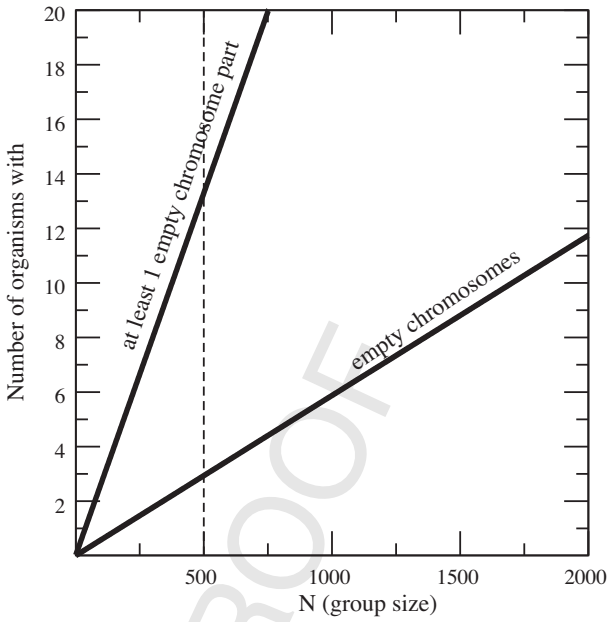

Fig. 2. Visual analysis of the viability of the organisms generated by the random breakage model as a function of N, the group size. Two solid lines are shown, one for the linear growth of the number of organisms with empty chromosomes (organisms such that $L_g > G$) that are produced by the random breakage model as the group size N increases and another one for the linear growth of the number of organisms with at least one empty main chromosome part (organisms such that $3L_g > G$; only the two telomeres and one centromere are considered as main parts). As a guide to the eye, a vertical line with $N = 500$ (dashed line) is included. The solid curves were obtained by averaging the results of simulating the random breakage model over 10^6 replicas. All the parameters of the model were taken from Fig. 2(b) of Solé (2010) except N.

2.2. Independence Between Chromosome Number and Genome Size in Well-formed Chromosomes

In most major groups of organisms examined by Ferrer-i-Cancho and Forns (2009), a significant correlation between genome size and chromosome number has been found (Table 1). Genome size tends to decrease as the number of chromosomes increases in angiosperm plants and jawless fishes while it tends to increase as chromosome number increases in many other groups of species: fungi, gymnosperm plants, insects, reptiles, mammals, ray-finned fishes and amphibians (Table 1). The assumption of independence between genome size and chromosome number made by the random breakage is only valid for birds and cartilaginous fishes according to Table 1. From the biological perspective, it has been suggested that the negative correlation between G and L_g in angiosperms would result from a trade-off between recombination mechanisms (Vinogradov, 2001).

2.3. The Dependency Between Mean Chromosome Length and Chromosome Number in the Random Breakage Model

Firstly, it will argued that a power law with a $\beta = -1$ exponent and an additive constant, i.e.

$$E[Lc|Lg] = \frac{a}{Lg} + d, \quad (7)$$

where a is a proportionality constant and d is an additive constant and $L_g > 0$, is not supported by the major groups of organisms

Please cite this article in press as: Baixeries, J., et al., Random models of Menzerath–Altmann law in genomes. BioSystems (2011), doi:10.1016/j.biosystems.2011.11.010

Table 1
Summary of results on the dependency between L_c (mean length of chromosomes) and L_g (chromosome number) and the dependency between G (genome size) and L_g. *Yes* is used to indicate statistically significant correlations at a significant level of 0.05. + and − are used to indicate the sign of the correlation. (*) is used to indicate results borrowed from Ferrer-i-Cancho and Forns (2009). The significant and non-significant correlations remain if the updated dataset in Hernández-Fernández et al. (2011) is used. (**) is used to indicate results borrowed from Hernández-Fernández et al. (2011). The results on the correlation between G and L_g in angiosperm plants are supported independently by the pioneering research by Vinogradov (2001). The hypothesis of linearity between G and L_g was rejected by means of a non-parametric linearity test (Hernández-Fernández et al., 2011).

Group	Correlation between L_c and L_g (*)	Correlation between G and L_g (**)	Non-linear dependency between G and L_g (**)
Fungi	Yes −	Yes +	Yes
Angiosperm plants	Yes −	Yes −	Yes
Gymnosperm plants		Yes +	Yes
Insects	Yes −	Yes +	Yes
Reptiles	Yes −	Yes +	Yes
Birds	Yes −		Yes
Mammals	Yes −	Yes +	Yes
Cartilaginous fishes	Yes −		Yes
Jawless fishes	Yes −	Yes −	Yes
Ray-finned fishes		Yes +	Yes
Amphibians	Yes +	Yes +	Yes

considered by Ferrer-i-Cancho and Forns (2009) and Hernández-Fernández et al. (2011). To see it, notice that the fact that $L_c = G/L_g$ means that Eq. (7) is equivalent to

$$E[G|L_g] = a + dL_g, \tag{8}$$

namely that $E[G|L_g]$ is a linear function of L_g. Interestingly, the linear dependency between genome size and chromosome number that Eq. (8) defines has been rejected by means of a non-parametric linearity test (Table 1), which indirectly rejects Eq. (7). Further support for the non-linear dependency between G and L_g can be obtained assuming that G is exactly the sum of all chromosome sizes, i.e. (Solé, 2010)

$$G = \sum_{i=1}^{L_g} g_i, \tag{9}$$

where g_i is the size in base pairs of the ith chromosome. The definition in Eq. (9) implies that if $L_g = 0$ was reached then $E[G|L_g] = 0$, which gives $a = 0$ in Eq. (8). Notice that $a = 0$ implies $d \neq 0$ as a genome cannot be empty when $L_g > 0$. Two predictions of $a = 0$ on Eqs. (7) and (8), $E[L_c|L_g] = d$, i.e. no significant correlation between L_c and L_g, and $E[G|L_g] = dL_g$ with $d \neq 0$, i.e. a significant correlation between G and L_g, are not found in any major group of organisms except in gymnosperm plants and ray-finned fishes (Table 1). The fact that this alternative non-parametric linearity test is not able to reject linearity for these two groups of organisms is not a problem for our arguments: Eq. (7) can be discarded because Eq. (8) is rejected or because a significant negative correlation between L_c and L_g is not found as in gymnosperm plants and ray-finned fishes (Table 1).

Secondly, a general class of random models, of which the random breakage model is a particular case, will be presented. It will be shown that the models of this class obey Eq. (7) with $d = 0$ and a being determined solely by the distribution of G. For the random breakage model above it will be shown that

$$a = \frac{G^M(G^M - 1) - (G^m + 1)G^m}{2(G^M - G^m - 1)}. \tag{10}$$

Consider an extended random model that generates the information about an organism in two stages. It is assumed that $G, L_g \geq 1$. In the first stage, the values of G and L_g for a group of N organisms through the following procedure:

- G is chosen at random from a certain distribution A.
- L_g is chosen at random from a certain distribution B and independently from G.
- Given G and L_g, L_c is computed through $L_c = G/L_g$.

In the second stage, the length of each of the L_g chromosomes is generated following a certain procedure C.

Uniform distributions were selected for distributions A and B in the random breakage model used to explain correlations consistent with Menzerath–Altmann law in genomes (Solé, 2010). Reciprocal translocation and random breakage are examples of mechanisms that have been considered for procedure C (De et al., 2001; Li et al., in press; Sankoff and Ferretti, 1996; Solé, 2010). However, the distribution B and procedure C are irrelevant for $E[L_c|L_g]$ in these general class of random models. In all the calculations that follow, true independence between G and L_g is assumed. This means that we assume that chromosomes of length zero or chromosome with empty main parts are not discarded. The following general theorem will be used to derive the exact mathematical form of $E[L_c|L_g]$ in this general class of random models.

Theorem 1. *Two random natural numbers X and Y, such that $X > 0$, are independent if and only if $Z = Y/X$ satisfies $E[Z|X] = E[Y]/X$.*

Proof. By definition of independence between X and Y,

$$p(Y = y|X = x) = p(X = x) \tag{11}$$

for any x and y. Multiplying by $z = y/x$ on both sides of the previous equality, it is obtained

$$p(Y = y|X = x)z = p(X = x)\frac{y}{x} \tag{12}$$

for any x and y. Applying the fact that $Y = y$ is equivalent to $Y/X = y/x$, Eq. (12) leads to

$$p(Z = z|X = x)z = p(X = x)\frac{y}{x} \tag{13}$$

for any x and y thanks to the definitions $Z = X/Y$ and $z = y/x$.

Summing over y on both sides of the equality of Eq. (13) yields

$$\sum_y p(Z = z|X = x)z = \frac{1}{x}\sum_y p(Y = y)y \tag{14}$$

for any x. By the definition of expectation and conditional expectation and $z = y/x$, Eq. (14) is transformed into

$$E[Z|X = x] = \frac{E[Y]}{x} \tag{15}$$

for any x, which finally gives $E[Z|X] = E[Y]/X$ as we wanted to prove.

For the general class of random models above, the following corollary shows that $E[L_c|L_g]$ is a power law whose exact form is determined solely by distribution A (distribution B and procedure C are irrelevant) and that the exponent of the law does not depend on A.

Corollary 1. *The general class of random models above is equivalent to the class of models yielding*

$$E[L_c|L_g] = aL_g^{\beta}, \tag{16}$$

with $a = E[G]$ and $\beta = -1$.

Proof. Straightforward from Theorem 1 taking X as L_g, Y as G, Z as L_c and $a = E[G]$.

The exact form of $E[L_c|L_g]$ for the random breakage model (Solé, 2010) will be derived applying some elementary results on uniformly distributed random numbers that will be obtained first.

Please cite this article in press as: Baixeries, J., et al., Random models of Menzerath–Altmann law in genomes. BioSystems (2011), doi:10.1016/j.biosystems.2011.11.010

Theorem 2. *Let X be a uniformly distributed integer random variable within the interval* (x^m, x^M)*, i.e.* $p(X=x) = 1/(x^M - x^m - 1)$ *if* $x \in (x^m, x^M)$ *and* $p(X=x) = 0$ *otherwise.*

The expectation of X is

$$E[X] = \frac{x^M(x^M-1) - (x^m+1)x^m}{2(x^M - x^m - 1)}. \quad (17)$$

Proof.

$$E[X] = \sum_{x=x^m+1}^{x^M-1} p(X=x)x = p(X=x)\left(\sum_{x=1}^{x^M-1} x - \sum_{x=1}^{x^M} x\right)$$

$$= \frac{x^M(x^M-1) - (x^m+1)x^m}{2(x^M - x^m - 1)}. \quad (18)$$

Corollary 2. *The expection of G and* L_c *in the random breakage model* (Solé, 2010) *are respectively*

$$E[G] = \frac{G^M(G^M-1) - (G^m+1)G^m}{2(G^M - G^m - 1)} \quad (19)$$

$$E[L_g] = \frac{L_g^M(L_g^M-1) - (L_g^m+1)L_g^M}{2(L_g^M - L_g^m - 1)}. \quad (20)$$

Proof. Applying Theorem 2 for G with $x_m = G^m$ and $x^M = G^M$ gives Eq. (19). Applying Theorem 2 for L_g with $x^m = L_g^m$ and $x^M = L_g^M$ gives Eq. (20).

Corollary 3. *In the random breakage model* (Solé, 2010)*, the expectation of* L_c *when* L_g *is given is* $E[L_c|L_g] = a/L_g$ *with a defined as in* Eq. (10).

Proof. Straightforward applying Corollaries 1 and 2.

3. The Effect of Viability on the Dependency Between the Size of the Parts and the Size of the Whole

Corollary 1 states that $E[L_c|L_g] \sim 1/L_g$ can only be achieved by independence between G and L_g. Next the effect of a particular viability concern on the shape of $E[L_c|L_g]$ will be studied in the simple random breakage model (Solé, 2010). As expected from Corollary 1, the new $E[L_c|L_g]$ deviates from a/L_g.

For simplicity, let us assume that the main three parts, i.e. two telomeres and the centromere, must have of least length $k \geq 0$ each, i.e. c^m, t_1^m, $t_2^m = k$ for an organism to be viable (we assume that the remainder of the parts could be empty). Thus a viable organism needs $3kL_g \leq G$ (when $k=0$ no viability constraint is imposed). Under these constraints, the next theorem describes the expected mean chromosome length as a function chromosome number.

Theorem 3. *Consider that the chromosome number* L_g $(L_g > 0)$ *and the genome size G are natural numbers. Consider also that* L_g *is given and that then G must be generated. If a chromosome has three main parts that each must have length equal or greater than k (with* $k \geq 0$*) in order to be viable (the remainder of the parts can be empty) it follows that*

(1) *If* $3kL_g \leq G^m + 1$ *then all organisms of* L_g *chromosomes are viable, i.e.* $E[L_c|L_g] = a/L_g$*, with a defined as in* Eq. (10).
(2) *If* $3kL_g \geq G^M$ *then no organism of* L_g *chromosomes is viable, i.e.* $E[L_c|L_g]$ *is undefined.*
(3) *If* $G^m + 1 < 3kL_g < G^M$ *then some organisms of* L_g *chromosomes are not viable and*

$$E[L_c|L_g] = \frac{G^M(G^M-1) - 3kL_g(3kL_g-1)}{2L_g(G^M - 3kL_g)}. \quad (21)$$

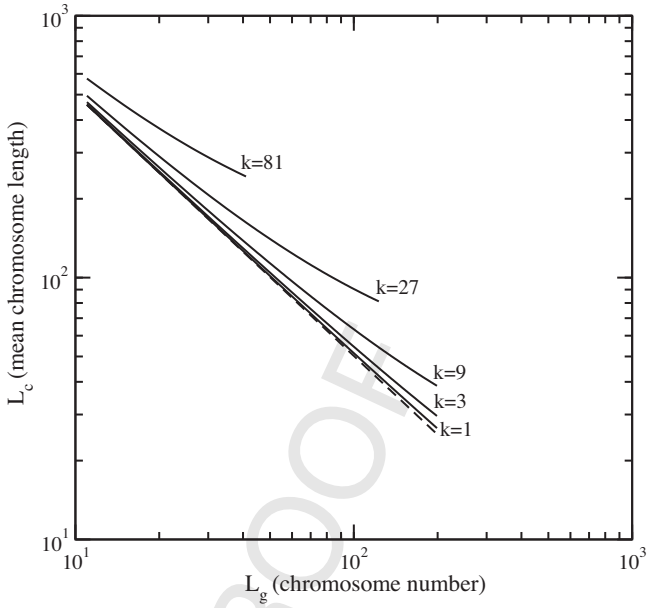

Fig. 3. A comparison of the expected value of L_c given L_g as a function of the viability factor k in the extended random breakage model. In this extension, an organism is not viable unless $3kL_g \leq G$, where L_g is the chromosome number of the organism and G is its genome size. According to this criterion, all non-viable organisms are removed. Dashed line is used for $k=0$ (no viability constraints) while solid line is used for $k>0$. Parameter setting (b) from Solé (2010), i.e. $G^m = 50$, $G^M = 10{,}000$, $L_g^m = 10$ and $L_g^M = 200$, was used. Theoretical expectations for the mean chromosome length for different values of k were obtained applying Theorem 3.

Proof. The proof of (1) follows from Corollary 3. The proof of (2) is straightforward as no viable organism can be generated in that parameter setting. As for the proof of (3), consider the value of G of a viable organism follows a uniform distribution within the open interval $(3kL_g - 1, G^M)$. That is, the distribution keeps being uniform as in the original random breakage model (Solé, 2010) when viability constraints are taken into account but with a left-truncation introduced by viability. By Theorem 2 with $x^m = 3kL_g - 1$ and $x^M = G^M$, it is obtained

$$E[G|L_g] = \frac{G^M(G^M-1) - 3kL_g(3kL_g-1)}{2(G^M - 3kL_g)}. \quad (22)$$

Knowing that $E[L_c|L_g] = E[G/L_c|L_g] = E[G|L_g]/L_g$, it is finally obtained

$$E[L_c|L_g] = \frac{G^M(G^M-1) - 5kL_g(5kL_g-1)}{2L_g(G^M - 5kL_g)} \quad (23)$$

as we wanted to prove.

Fig. 3 shows $E[L_c|L_g]$ for different values of k employing parameter setting (b) of (Solé, 2010). It can be seen that k introduces a dependency between G and L_g that leads to a deviation between $E[L_c|L_g]$ and a power-law with -1 exponent (a straight line in double logarithmic scale with -1 slope) that is expected from independence between G and L_g. The exact dependency between G and L_g that was chosen for Theorem 3 was used simply to illustrate the consequences of breaking the independence that has been assumed and defended (Solé, 2010). More realistic dependencies between G and L_g could be defined.

4. Discussion

There is a serious design flaw in the random breakage model: the explicit assumption that genome size and chromosome number are and must be independent. Referring to the models by Sankoff and Ferretti (1996) and De et al. (2001), it has been stated that

Please cite this article in press as: Baixeries, J., et al., Random models of Menzerath–Altmann law in genomes. BioSystems (2011), doi:10.1016/j.biosystems.2011.11.010

"The good fit obtained by these theoretical approaches and the current knowledge on the fluid nature of chromosomal rearrangements through time rule against any special multiscale link between genome-level and chromosome-level patterns" (Solé, 2010).

This argument is problematic for many reasons:

- In these models (Sankoff and Ferretti, 1996; De et al., 2001), chromosomes lengths are generated from a constant genome size and a constant chromosome number that are borrowed from a target species. These numbers are kept constant by the dynamical rules of the models. Therefore, these models constitute no evidence against "multiscale links".
- As we have discussed in the present article, a multiscale link is determined by the kind of constraints on chromosome structure (e.g., non-empty chromosome parts) that models of chromosome evolution have taken into account (Sankoff and Ferretti, 1996; De et al., 2001).
- Statistical studies have unraveled significant dependencies between genome size and chromosome number (Table 1; see also Trivers et al., 2004). It is possible to predict, for a given species, chromosome sizes by chromosome number, and furthermore, given either genome size or average chromosome length it is possible to predict the size range of all chromosomes of that species (Li et al., in press).
- Experiments indicate that "upper and lower tolerance limits for chromosome size are apparently determined by the genome size, chromosome number and karyotype structure of a given species" (see Schubert (2007) and references therein).
- It is has been argued that multiscale links could result from the interplay between opposite biological forces (Schubert, 2007 and references therein), e.g., trade-offs between recombination mechanisms.
- It has been shown here that the independence assumption of the random breakage model leads to a "power-law" dependency between mean chromosome length and chromosome number (a particular case of Menzerath–Altmann law, Eq. (1) with $b = -1$ and $c = 0$) that is not supported by actual genome data.

Any model is a simplification of reality and many models are only focused on a few statistical properties or simply one target statistical property as the random breakage model. The key is not making simplifications that do not let a model reproduce one of its target statistical properties. In the simple random breakage model (Solé, 2010), the target distribution is the dependency between L_c and L_g. Unfortunately, the independence between G and L_g defended in the random breakage model leads to a power-law dependency between L_c and L_g with a -1 exponent that has been rejected (recall Section 2.3). The only two taxonomic groups examined to support the random breakage model, plants and mammals (Solé, 2010), are among the groups for which that power-law provides an insufficient fit.

There are still more aspects of the random breakage model (Solé, 2010) that need to be revised. Firstly, the model generates chromosome lengths that are produced by selecting break points in the genome sequence uniformly at random and independently. Random breakage is known to be a suitable model in extremal conditions, e.g., the fragmentation of DNA induced by high dosages of ^{7}Li radiation (Fuquan et al., 2005). The idea that a genome can break at any point in normal circumstances has been refuted by independent computational analyses of genomes (e.g., Peng et al., 2006; Hinsch and Hannenhalli, 2006; Ruiz-Herrera et al., 2006). The conclusion of these studies is that the breaks are found in "fragile regions" or "hotspots" (Becker and Lenhard, 2007). Besides, random breakage produces chromosome lengths that follow the broken stick distribution (Smart, 1976). A recent study shows that the gamma distribution gives the best fit among various candidates (not including the broken stick distribution) to actual chromosome lengths (Li et al., in press). The quality of the fit a broken stick distribution to actual chromosome lengths should be evaluated and compared to that of a gamma distribution. Obtaining a sufficiently good fit to actual chromosome lengths with a broken stick distribution is a serious challenge for the random breakage model (Solé, 2010) because random breakage or fragmentation is not used as a true model but as a control for more realistic models of chromosome length evolution (Sankoff and Ferretti, 1996; De et al., 2001) and, in general, these more realistic models yield a better fit to actual data than random breakage. This challenge cannot be obviated since, at present, the actual statistical pattern of genomes that the random breakage model (Solé, 2010) is able to reproduce to a sufficient degree is not known. Besides, it should also be checked rigorously if genome sizes and chromosome number are uniformly distributed as the random breakage model assumes. Uniformity for genomes sizes is clearly not the case in general. The genome size of eukaryotes varies over five orders of magnitude but the distribution is skewed towards small values (Oliver et al., 2007). The skewness of the distribution of genome sizes has been demonstrated within angiosperm plants but the direction of the skewness (towards small values or towards large values) varies according to the subgroup under consideration (Vinogradov, 2001). It has been argued that eukaryotic genome sizes might evolve in a stochastically proportionate manner, which necessarily produces far more small genomes than large genomes, even in the absence of selection against large genomes (Oliver et al., 2007). Actual genome sizes at the level of all eukaryotes are consistent with a log-normal distribution (Oliver et al., 2007), which is predicted by proportional evolution after sufficiently long periods of time (Lewontin and Cohen, 1969).

In sum, the random breakage model and other models within the same class cannot trivially explain the dependency between mean chromosome length and chromosome number that is found in genomes. These models should be extended with realistic genome-chromosome dependencies in order to provide a satisfactory fit to actual genome data. At present, these models raise no serious objection to the possibility of a connection between linguistic systems and genomes through Menzerath–Altmann law.

Role of the Funding Source

The funder had not role in the study, design, collection, analysis and interpretation of data, the writing of the report and the decision to submit the paper for publication.

Acknowledgements

This article is dedicated to the memory of A. Hernández Carpio. We are grateful to N. Forns for a careful review and D. Menacho for helpful discussions. This work was supported by the project SESAAME-BAR (TIN2008-06582-C03-01) of the Spanish Ministry of Science and Innovation (JB and RFC).

References

Altmann, G., 1980. Prolegomena to Menzerath's law. Glottometrika 2, 1–10.
Becker, T.S., Lenhard, B., 2007. The random versus fragile breakage models of chromosome evolution: a matter of resolution. Molecular Genetics and Genomics 278, 487–491.
Boroda, M.G., Altmann, G., 1991. Menzerath's law in musical texts. Musikometrica 3, 1–13.
Cramer, I., 2005. The parameters of the Altmann–Menzerath law. Journal of Quantitative Linguistics 12 (1), 41–52.

De, A., Ferguson, M., Sindi, S., Durrett, R., 2001. The equilibrium distribution for a generalized Sankoff–Ferretti model accurately predicts chromosome size distribution in a wide variety of species. Journal of Applied Probability 38, 324–334.
Ferrer-i-Cancho, R., Forns, N., 2009. The self-organization of genomes. Complexity 15 (5), 34–36.
Fuquan, K., Kui, Z., Yong, Z., Tianguang, C., Meinan, N., Li, S., Minghui, C., Yizhong, Z., 2005. Analysis of length distribution of short DNA fragments induced by ^{7}Li ions using the random-breakage model. Chinese Science Bulletin 50 (9), 841–844.
Hernández-Fernández, A., Baixeries, J., Forns, N., Ferrer-i-Cancho, R., 2011. Size of the whole versus number of parts in genomes. Entropy 13 (8), 1465–1480, doi:10.3390/e13081465.
Hinsch, H., Hannenhalli, S., 2006. Recurring genomic breaks in independent lineages support genomic fragility. BMC Evolutionary Biology 6, 90.
Lewontin, R.C., Cohen, D., 1969. On population growth in a randomly varying environment. Proceedings of the National Academy of Sciences of the United States of America 62, 1056–1060.
Li, X., Zhu, C., Lin, Z., Wu, Y., Zhang, D., Bai, G., Song, W., Ma, J., Muehlbauer, G.J., Scalon, M.J., Zhang, M., Yu, J. Chromosome size in diploid eukaryotic species centers on the average length with a conserved boundary. Molecular Biology and Evolution, in press.
Luisi, P.L., 2006. Chapter 11: approaches to the minimal cell. In: The Emergence of Live. From Chemical Origins to Synthetic Biology. Cambridge University Press, Cambridge, pp. 242–267.
Oliver, M.J., Petrov, D., Ackerly, D., Falkowski, P., Schofield, O.M., 2007. The mode and tempo of genome size evolution in eukaryotes. Genome Research 17, 594–601.
Peng, Q., Pevzner, P.A., Tesler, G., 2006. The fragile breakage versus random breakage models of chromosome evolution. PLoS Computational Biology 2, e14.
Ruiz-Herrera, A., Castresana, J., Robinson, T.J., 2006. Is mammalian chromosomal evolution driven by regions of genome fragility? Genome Biology 7, R115.
Sankoff, D., Ferretti, V., 1996. Karyotype distributions in a stochastic model of reciprocal translocation. Genome Research 6, 1–9.
Schubert, I., 2007. Chromosome evolution. Current Opinion in Plant Biology 10, 109–115.
Smart, J.S., 1976. Statistical tests of the broken-stick model of species-abundance relations. Journal of Theoretical Biology 59 (1), 127–139.
Solé, R.V., 2010. Genome size, self-organization and DNA's dark matter. Complexity 16 (1), 20–23.
Teupenhayn, R., Altmann, G., 1984. Clause length and Menzerath's law. Glottometrika 6, 127–138.
Trivers, R., Burt, A., Palestis, B.G., 2004. B chromosomes and genome size in flowering plants. Genome 47 (1), 1–8.
Vinogradov, A.E., 2001. Mirrored genome size distributions in monocot and dicot plants. Acta Biotheoretica 49, 43–51.
Wilde, J., Schwibbe, M.H., 1989. Organizationsformen von Erbinformation Im Hinblick auf die Menzerathsche Regel. In: Altmann, G., Schwibbe, M.H. (Eds.), Das Menzerathsche Gesetz in informationsverarbeitenden Systemen. Olms, Hildesheim, pp. 92–107.

Please cite this article in press as: Baixeries, J., et al., Random models of Menzerath–Altmann law in genomes. BioSystems (2011), doi:10.1016/j.biosystems.2011.11.010