Comparing the Statistical Fate of Paralogous and Orthologous Sequences

Florian Massip,[1,2,3] Michael Sheinman,[4,2] Sophie Schbath,[1] and Peter F. Arndt[2]

[1]*INRA, UR1404 Mathématique Informatique Appliquées du Génome à l'Environnement, Jouy-en Josas, France*
[2]*Max Planck Institute for Molecular Genetic, Berlin, Germany*
[3]*Laboratoire Biométrie et Biologie Évolutive, CNRS, UMR5558, Université de Lyon, Université Lyon 1, Villeurbanne, France*
[4]*Department of Biology,Faculty of Science, Utrecht University, Utrecht, the Netherlands*

Since several decades, sequence alignment is a widely used tool in bioinformatics. For instance, finding homologous sequences with known function in large databases is used to get insight into the function of non-annotated genomic regions. Very efficient tools, like BLAST have been developed to identify and rank possible homologous sequences. To estimate the significance of the homology, the ranking of alignment scores takes a background model for random sequences into account. Using this model one can estimate the probability to find two exactly matching subsequences by chance in two unrelated sequences. The corresponding probability for two homologous sequences is much higher allowing to identify them. Here we focus on the distribution of lengths of exact sequence matches in protein coding regions pairs of evolutionary distant genomes. We show that this distribution exhibits a power-law tail with exponent $\alpha = -5$. Developing a simple model of sequence evolution by substitutions and segmental duplications, we show analytically that paralogous and orthologous gene pairs contribute differently to this distribution. Our model explains the differences observed in the comparison of coding and non-coding parts of genomes, thus providing with a better understanding of statistical properties of genomic sequences and their evolution.

I. INTRODUCTION

One of the first and most celebrated bioinformatic tools is sequence alignment [1, 17, 24]. Even today, the development of algorithms that are able to search for similar sequences of a query sequence in a huge database is still an active research field. In this matter, one needs to be able to distinguish sequence alignments that are due to a biological relatedness of two sequences from those which occur by random. Let us, for simplicity, disregard mismatching nucleotides and insertions and deletions (indels or gaps) in an alignment and only consider so-called maximal exact matches, i.e. sequences that are 100% identical and cannot be extended on both ends. In this case, the length distribution of matches is equivalent to the score distribution and can easily be calculated for an alignment of two random sequences where each nucleotide represents an i.i.d. random variable. We expect the number of matches to be distributed according to a geometric distribution, such that the number, $M(r)$, of exact maximal matches of length r is given by

$$M(r) = p^r(1-p)^2 L_A L_B, \tag{1}$$

where L_A and L_B are the lengths of the two genomes, p^r is the probability that r nucleotides match, and $(1-p)^2$ is the probability that a match is flanked by two mismatches. Here $p = \sum_\alpha f_A(\alpha) f_B(\alpha)$, where $f_X(\alpha)$ is the frequency of nucleotide α in the genome of species X and the sum is taken over all nucleotides. Thus, the number of matches for a given length r is expected to decrease very fast

as the length r increases, and for generic random genomes of hundreds of Mbp, one does not find any match longer than 25 bp.

For long matches, real genomes strongly violate the prediction of Eq. (1) due to the evolutionary relationships between and within genomes [21]. Comparing the genomes of recently diverged species, one finds regions in the two genomes that have not acquired even a single substitution. In the following, substitution refers to any genomic change which would disrupt a 100% identical match (as for instance a nucleotide exchange, an insertion, or a deletion). As the divergence time between the two species increases, such matches will get smaller very fast and most remaining long matches will be found either in exons or in ultraconserved elements [2], that both evolve under purifying selection.

Computing the match length distribution (MLD) from the comparison of human and mouse genomes, we thus expect to observe much longer exact matches than in a comparison of two random sequences of the same lengths. The observed MLD for exons and non-coding sequences in the human and mouse genomes is shown in Fig. 1. At the left end of the distribution, i.e. for $r < 25$ bp, the distribution is dominated by random matches, as described by Eq. (1). As expected, this MLD deviates from the random model for matches longer than 25 bp.

Interestingly, in this asymptotic regime, the MLD exhibits a power-law tail $M(r) \sim r^{\alpha}$ (identified as a straight line in the double logarithmic plot), where the exponent α is close to -5 for exonic sequences. This is in contrast to the MLD of non-coding sequences, where the exponent α is close to -4 [9, 10, 16]. This property appears to be impressively reproducible in the comparison of various pairs of species (see Fig S1). The value of the exponent α was calculated using the maximum likelihood estimator. To assess the robustness of this estimator, we also performed a bootstrap analysis that showed good agreement with the estimated value of α, see Section V in the supplementary material. Discrepancies from the of power-law behavior can be observed for very long matches, since such matches are scarce and random noise distorts the distributions.

It is tempting to speculate that this peculiar behavior of exonic sequences is a direct consequence of their coding function and might reflect structural or other constrains of proteins. In the following, we demonstrate that this distribution can be explained by a simple evolutionary model that takes into account the generation of paralogous sequences (due to segmental duplications) and orthologous sequences (due to speciation) [6]. Further, we assume that paralogous and orthologous matches are subsequently broken down by random substitutions. This dynamics can be modelled by a well-known stick-breaking process [13, 26] introduced below. Since our model describes the existence of long matching sequence segments in two genomes, it also has to include selection. However we model selection in a minimal way, since we only assume that regional substitution rates are distributed, such that there are regions that evolve very slowly. Our model can therefore be viewed as a minimal model for evolution of functional sequences, which reproduces certain statistical features of their score distributions. We proceed now with the detailed introduction of the model.

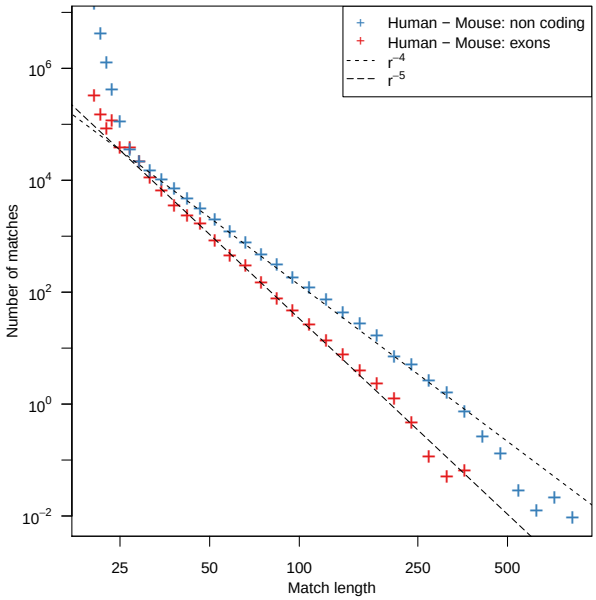

Figure 1: Two MLDs computed from the comparison of different subsets of the Human and the Mouse genomes. One MLD was computed from the comparison of the RepeatMasked non-coding part of both genomes (blue crosses) and the second is the result of the comparison of the coding part of both genomes (red crosses). Dashed lines represent power-law distributions with exponents $\alpha = -4$ and $\alpha = -5$. All data are represented using a logarithmic binning to reduce the sampling noise [18], see the Materials and Methods section in the Supplementary Material for more details.

II. THEORY

The Stick-Breaking Model — Before we turn to the detailed description of our model, let us shortly introduce some relevant results on random stick-breaking. Consider a stick of length K at time $t = 0$, which will be sequentially broken at random positions into a collection of smaller sticks. Breaks occur with rate μ per unit length. The distribution of stick lengths at time t, denoted by $m(r, t)$, follows the following integro-differential equation [15, 26]:

$$\frac{\partial}{\partial t} m(r,t) = -\mu\, r\, m(r,t) + 2\mu \int_r^\infty m(s,t) ds, \tag{2}$$

where the first term on the right hand side represents the loss of sticks of length r due to any break in the given stick and the second term represents the gain of sticks of length r from the disruption of longer sticks. Note that for any stick of length $s > r$, there are two possible positions at which a break would generate a stick of length r.

The initial state is one unbroken stick of length K, i.e. $m(r, 0) = \delta(K, r)$. The corresponding time-dependent solution is [26]:

$$m(r,\tau) = \begin{cases} [2\tau + \tau^2(K-r)]\exp(-\tau r) & \text{for } 0 < r < K, \\ \exp(-K\tau) & \text{for } r = K, \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where we define the rescaled time $\tau = \mu t$. Apart from the singularity at $r = K$, which accounts for the possibility that the stick is not even broken once, the distribution is dominated by an

exponential function, i.e. there are far more small sticks than long ones. The average stick length is given by $\bar{m}(\tau) = K/\tau$.

The Match Length Distribution of Evolving Sequences — The above stick-breaking process can be used to describe the break down of a long DNA match into several smaller ones by substitutions in either one of the two copies of the match. In a comparison of two species, A and B, long identical segments are the signature of homology relationships between the two sequences. These homologous sequences result either from the copy of the genetic material during the time of speciation, and are then orthologuous sequences (see the blue dashed line in Fig. 2) or due to segmental duplications in the ancestral genome, i.e. paralogous sequences (see the red dashed line in the same figure) [6].

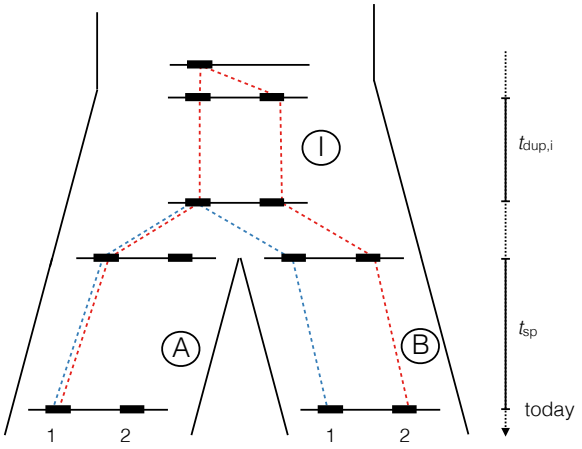

Figure 2: The different contributions to the match length distribution. Sequence 1 was duplicated in the ancestral species I. This duplication gives rise to two paralogous sequence pairs: sequence 1 in A with sequence B 2 in B (red dashed line), and sequence B 1 and sequence A 2. Sequence 1 in A is orthologous to sequence 1 in B (blue dashed line), and Sequence 2 in A is orthologous to sequence 2 in B. For clarity, we highlight only one pair only for each exemple on the figure.

The match length distribution (MLD) is then given by the integral

$$M(r) = \int_0^\infty N(\tau) m(r,\tau) d\tau \tag{4}$$

where $N(\tau)$ is the number of homologous sequences with divergence τ and $m(r,\tau)$ is given in Eq. (3), see also Ref.[16]. The divergence between a pair of orthologous sequences is the sum of two contributions $\tau = \mu_{A,i}\, t_{\rm sp} + \mu_{B,i}\, t_{\rm sp}$, where $t_{\rm sp}$ is the time since the two species diverged and i is an index for regions in the genomes. The regional mutation rates $\mu_{A,i}$ and $\mu_{B,i}$ in the two species are themselves distributed and assumed to be independent from each other. We therefore define N_{AB}

$$N_{AB}(\tau) = \int_0^\tau N_A(\tau_A)\, N_B(\tau - \tau_A)\, d\tau_A \tag{5}$$

where $N_A(\tau)$ (resp. $N_B(\tau)$) is the number of sequences with divergence τ from the last common ancestor I in species A (resp. B). However if the two regions are paralogous, the divergence τ is a sum of three independent contributions $\tau = \mu_{A,i}\, t_{\rm sp} + (\mu_{I,i} + \mu_{I,j})\, t_{\rm dup} + \mu_{B,j}\, t_{\rm sp}$ where $t_{\rm dup}$

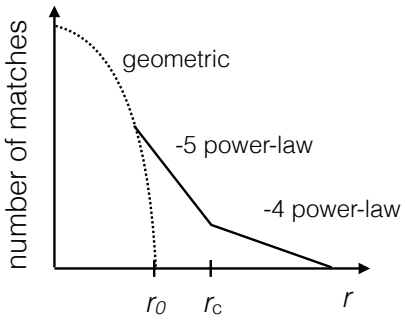

Figure 3: Schematic drawing of the match length distribution in a double logarithmic plot. The two regimes exhibiting a -4 and -5 power-law (continuous lines) are separated by a cross-over point. For very small match lengths the geometric distribution due to random matches, see Eq. (1), dominates (dotted line).

represent the time elapsed between the segmental duplication and the split of the two species. There are $N_{AIB}(\tau)$ paralogous sequences with divergence τ, with

$$N_{AIB}(\tau) = \int_0^{\tau} \int_0^{\tau-\tau_A} N_A(\tau_A)\, N_I(\tau - \tau_A - \tau_B)\, N_B(\tau_B)\, d\tau_B\, d\tau_A\,. \tag{6}$$

For our purposes we are not interested in the full functional form of the distributions in Eqs. (5) and (6) but only have to consider their behavior for small $\tau \to 0$, because long matches (and thus the tail of the distribution of the match length distribution $M(r)$) stem from homologous exons that exhibit a small divergence τ. A more general discussion about the functionnal form of the distrbution of pairwise distances can be found in ref.[23]. We therefore take the Taylor expansion of the distributions $N(\tau)$ around $\tau = 0$. Using Leibniz's formula to take the derivative under the integral sign [7] we find for orthologous exons (see details in the Supplementary Material)

$$N_{AB}(\tau) = N_A(0) N_B(0)\, \tau + \mathcal{O}(\tau^2)\,. \tag{7}$$

and subsequently the match length distribution [16]

$$\begin{aligned} M_{AB}(r) &= \int_0^{\infty} N_{AB}(\tau) m(r,\tau) d\tau \\ &= N_A(0) N_B(0) \frac{6K - 2r}{r^4} \\ &\sim N_A(0) N_B(0) \frac{6K}{r^4}\,, \end{aligned} \tag{8}$$

as K is much larger than r. In contrast, expanding Eq. (6) around $\tau = 0$, one can see that for paralogous pairs, the number of regions with divergence τ increases as τ^2 in the small τ limit (see details in the Supplementary Material)

$$N_{AIB}(\tau) = \frac{1}{2} N_A(0) N_I(0) N_B(0)\, \tau^2 + \mathcal{O}(\tau^3)\,. \tag{9}$$

Therefore the match length distribution exhibits a power-law tail with exponent $\alpha = -5$:

$$\begin{aligned} M_{AIB}(r) &= \int_0^\infty N_{AIB}(\tau) m(r,\tau) d\tau \\ &= N_A(0) N_I(0) N_B(0) \frac{12K - 6r}{r^5} \\ &\sim N_A(0) N_I(0) N_B(0) \frac{12K}{r^5}\,, \end{aligned} \tag{10}$$

as K is much larger than r.

Depending on the number of orthologous sequences Q_{ortholog} and paralogous sequences Q_{paralog}, we will be able to distinguish two regimes: one where the MLD follows an $\alpha = -4$ power-law and one where it follows an $\alpha = -5$ power-law. From Eqs. (8) and (10), it is straightforward to find that the cross-over point r_c between those regimes (see Fig. 3) is at

$$r_c = 2N_I(0)\,. \tag{11}$$

Recall that $N_I(0)$ is defined as the number of paralogous segments, at the time of the split, that have not mutated even a single time since the duplication event. Thus, this term is proportional to the ratio of the duplication rate over the mutation rate. If $N_I(0) \gg 10$, there are significantly more paralogous sequences compared to orthologous ones and the cross-over value, r_c, is large. Then, only the $\alpha = -5$ power-law tail will be observed. On the other hand, if $N_I(0) \ll 10$, then the crossing point r_c is expected to be smaller than 20 such that the $\alpha = -5$ power-law only holds for lengths were the distribution is anyway dominated by random matches. In contrast to previous models[16], this model does take into account the contribution of paralogous sequences, and can explain both power-law behaviors and therefore predicts the crossing point between the two regimes.

III. NUMERICAL VALIDATION

Our theoretical considerations predict a complex behavior of the match length distribution under the described evolutionary dynamics. The key ingredients are segmental duplications, generating paralogous sequences in an ancestral genome, and point mutations, that break identical pairs of paralogous and orthologous sequences of the two genomes after speciation into smaller pieces. To illustrate our theoretical predictions concerning the two power-laws, as well as the existence of the cross-over point r_c, we simulated the evolution of sequences according to the discussed scenario.

We describe the evolution of a genome of length L according to two simple processes, point mutations and segmental duplications. Point mutations exchange one base pair by another one and occur with rate μ per bp and unit of time. Note that to mimic the existence of regions under different degrees of selective pressure, we allow for regional differences of the point mutation rates. Segmental duplications copy a contiguous segment of K nucleotides to a new position where it replaces the same amount of nucleotides, such that the total length of the genome stays constant. Segmental duplications occur with rate λ per bp.

The evolutionary scenario of our simulation has two stages (see Fig. 2). At time $t = 0$, we generate a random iid sequence $\mathcal{S}$. During a time t_0, this sequence evolves according to the two described processes. In this first stage, the mutation rate is the same for all positions. At the end of this stage, the sequence represents the common ancestral genome of two species. At the beginning of the second stage, we copy the entire sequence of the common ancestor to generate

the genomes of the two species A and B. These sequences are then subdivided into M continuous regions of equal length. In each such region j, the point mutation rates $\mu_{A,j}$ (resp. $\mu_{B,j}$) are the same for all sites i and are drawn from an exponential distribution with mean μ (i.e. the point mutation rate during the first stage). The exponential distribution stipulates the least information under the given constraints. For more details about the simulation procedure, see Section V in the Supplementary Material.

We show the result of of the comparison of simulated sequences on the left panel of Fig. 4. One can clearly identify a power-law tail in the match length distribution, which for match length $20 < r < 100$ has an exponent $\alpha = -5$ and an exponent $\alpha = -4$ for longer matches $r > 100$. For simulated sequences, we can also easily classify homologous sequences into orthologous and paralogous sequences (while for natural sequences, paralogs and orthologs are not as easy to distinguish due to genomic rearrangements). Thus, we can also compute the MLD obtained from the comparison of paralogs, and the MLD obtained from the comparison of orthologs for simulated sequences. We show these two different distributions on the right panel of Fig. 4, where one can clearly observe that orthologous sequences generate an $\alpha = -4$ power-law distribution while paralogous matches generate an $\alpha = -5$ power-law distribution. In such a plot, one can further easily identify the crossing point as the value of r for which we obtain more matches from the comparison of orthologs than from the comparison of paralogs.

In the previous section the value of this crossing point between the two regimes was predicted to be $r_c = 2N_I(0)$ see Eq. (11), where $N_I(0)$ is the number of paralogous segments with a divergence $\tau = 0$ just before the species split. In our simulation procedure, $N_I(0)$ is simply the number of sequences that have been duplicated but that have not been mutated yet at the splitting time $t = t_0$. This number is known to be $N_I(0) = \frac{\lambda L}{\mu K}$ [15]. In our simulations, we have chosen $\lambda = 0.05$, $\mu = 1$, $K = 1000$ and $L = 10^7$ and therefore the crossing over is predicted to happen at $r_c = 100$, which is in good agreement with our observations in Fig. 4. Thus, the results of our simulations are in good agreement with our analytical predictions.

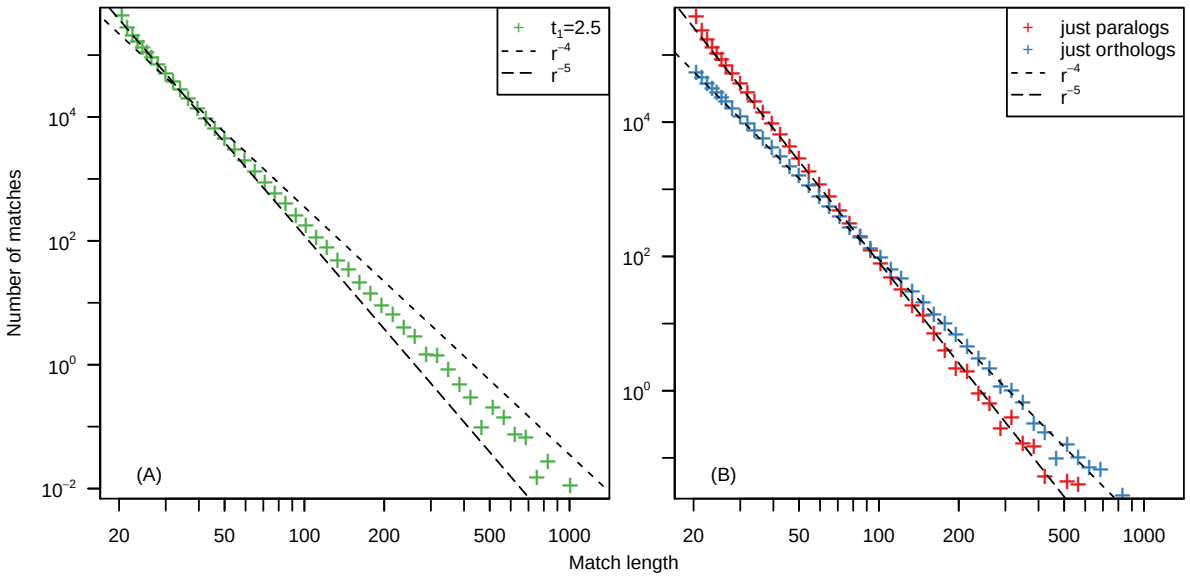

Figure 4: The MLD computed from 1000 simulated sequences according to the procedure described in the main text. Data are represented using the logarithmic binning. On the left panel, we show the MLD generated computed from all possible matches, while on the right panel, we represent two different MLDs: one computed from paralogous matches only (red), and one computed from the orthologous matches only (blue). One can see that the two different MLDs crosses close to the expected crossing value $r_c = 100$.

IV. DISCUSSION

We developed a simple model that accounts for power-law tails in the length distribution of exact matches between two genomes. Our model assumes regional differences of the selective pressure such that the substitution rates in a region are drawn from a certain distribution. However for naturally evolving exons the selective pressure varies also on shorter length scales. For instance, some nucleotides for many codons can be synonymously substituted by another one, mostly at third codon positions. Therefore, these substitutions at 3rd codon positions occur with a higher rate than non-synonymous ones. Hence, exons are expected to break preferentially at positions $3n$, with $n \in \mathbb{N}$, such that the matches with 100% identity would have lengths $3n + 2$ with integer n. Classifying genomic matches according to the remainder which is left when dividing their length by 3, we observe an almost 10-fold over-representation of matches with length $3n + 2$ over matches of lengths $3n$ and $3n + 1$, see Fig. 5. This suggests that the match-breaking process is dominated by the synonymous mutation rate which does not seem to vary significantly along an exon, while it does vary from one exon to another.

Using the presented model, the puzzling observation of an $\alpha = -4$ power-law tail in the MLD in the comparison of the human and mouse genomes and a corresponding $\alpha = -5$ power-law tail in the comparison of their exomes can be explained. Although the sequences stem from the same species, the relative amount of paralogous to orthologous sequence segments is different in the two data sets, which subsequently leads to different cross-over point r_c. Because of the selective pressure on coding exons, the number of non-mutated paralogous sequences at the time of species divergence

$N_I(0)$ is higher (relative to the number of orthologous sequences) in the exonic dataset than in the non-coding dataset. Thus, the cross-over point in exomes r_c is larger than the longest observed match and only the $\alpha = -5$ can be observed.

The opposite is true for matches in the alignment of non-coding sequences. Quantitatively, in this set, paralogous sequences play a lesser role and therefore only the $\alpha = -4$ power-law is observed (see Fig. 1). This is surprising, as the duplication rate is thought to be roughly the same in the coding and non-coding parts of genomes. To confirm this paradoxical observation, we classified matches according to the uniqueness of their sequences in both genomes. Assuming that unique matches are more likely to be orthologous, this gives us a rough classification of homologs into orthologs and paralogs, although matches unique in both sequences can be paralogs. After the classification of all matches, our analysis made apparent that matches unique in both genomes dominate the MLD in the comparison of the non-coding parts of the genomes, while matches with several occurrences in either (or both) genomes dominate the distribution in the case of the comparison of exomes (see Fig. S2 in the Supplementary Material). Moreover, we computed the MLD from the set of non-unique matches of the non-coding part of genomes. In this comparison, the contribution of paralogs is expected to be much higher than in the full set. As expected, this MLD also exhibits an $\alpha = -5$ power-law (see Fig. S2 in the Supplementary Material), confirming that the relative contribution of orthologs and paralogs is responsible for the shape of the MLD. These differences in the proportion of paralogous sequences in the coding and non-coding DNA is likely due to the fact that paralogs are more often retained in coding sequences than in the non-coding part of genomes. Since there are much more non-coding sequences in both genomes, we also observe at least 10 times more matches in the comparison of non-coding sequences than in the comparison of exomes.

The presented model does not account for changes in the divergence rates after a duplication, a phenomenon which is well documented following a gene duplication[11, 19, 20, 22]. To assess the impact of this phenomenon on the MLD, we performed simulations where the two paralogous segments are assigned different and independent mutation rates. Interestingly, these simulations yield qualitatively similar results than the simpler model introduced above (see Fig. S3). This new condition does not affect the value of the number of paralogous sequences that have not diverged at the time of the split (i.e. the value of $N_I(0)$), and thus the shape of the distribution.

The model we present is very simple, and more realistic models of genome evolution include many more evolutionary processes [5]. For instance one could include a transition/transversion bias in the mutational process, variations of mutation rates in time, a codon usage bias or different rates of duplications within and between chromosomes. However, since in the end we just consider identical matching sequences (and do not differentiate between miss-matches due to transitions and transversions) and want to explain the power-law tail in the MLD, all these additional model details are not expected to affect the results.

In this paper, we demonstrate that on the genome-wide scale, the length distribution of identical homologous sequence segments in a comparative alignment is non-trivial and exhibits a power-law tail, and we propose a simple model able to explain such distributions. While paralogous sequences, which had been duplicated before the species diverged generate a power-law tail with exponent $\alpha = -5$, orthologous sequences generate a power-law tail with exponent $\alpha = -4$. Depending on the relative amount of paralogous to orthologous sequences there is a cross-over between these two power-law regimes. The exponent of the power-law tail in the comparative MLD can therefore be a litmus test for the abundance of paralogous relative to orthologous sequences, while it is usually

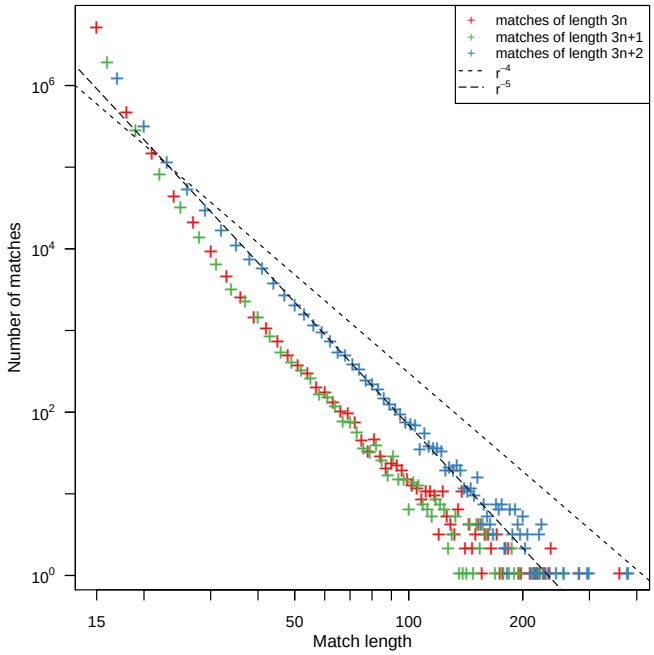

Figure 5: The MLD computed from the comparison of human and mouse exome, represented without logarithmic binning. 3 different colors are used to represent matches of length $3n$, $3n+1$ and $3n+2$. Dashed lines represent power-law distribution with exponent $\alpha = -4$ and $\alpha = -5$.

difficult to distinguish between orthologous and paralogous sequences using classical bioinformatic methods [4, 8, 25]. If paralogous sequences dominate (as for instance when comparing the exomes of human and mouse) the cross-over happens for long matches and the apparent exponent is equal to -5, otherwise it is equal to -4 (as for instance when comparing the non-coding part of human and mouse genomes) .

Our method is very easy to apply. In particular, it does not require that genomes are fully assembled as long as the continuous sequences are longer than 1kbp, comparable to the longest matches one expects. A natural extension of our method would be to apply it to sequences from metagenomic samples in order to assess relative amounts of paralogous and orthologous sequences. However, in this setting also horizontal gene transfer, which is common among prokaryotes and generates homologous sequence segments even between unrelated genomes, has to be included into the model. Our computational model can be easily be extended to take into account these and other more complex biological processes, using for instance already developed tools [5], which would allow to assess their impact on our results. The analysis of these models will be the subject of future work.

This study shows that even very simple models can often successfully be applied to seemingly very complex phenomena in biology. We were able to present a minimal model for the evolution of homologous sequences that includes effects due to segmental duplications and evolution under selective constrains — the two processes that are responsible for a power-law tail in the length distribution of identical matching sequences.

V. MATERIAL AND METHODS

Genomes — Unless otherwise stated, all genomes and their specific annotations (such as repeat elements and exons) were downloaded from the `ensembl` website [3] using the Perl API (version 80); the corresponding release of the Human genome is GRCh37. In all cases, we downloaded the RepeatMasked versions of genomes available in `ensembl` databases.

Computing MLDs — To find all identical matches (both in the case of self and comparative alignments), we used the `mummer` pipeline [14] (version 3.0), which allows to find all maximal exact matches between a "query" and a reference sequence using a computationally efficient suffix tree approach. For our analyses, we used the following options:

- `-maxmatch` such that `mummer` searches for all matches regardless of their uniqueness (by default, only matches unique in the query sequence are retrieved).
- `-n` that states that only 'A', 'T', 'C' and 'G' can match (any other character always results in a mismatch).
- `-b` such that `mummer` searches for matches on both strands. To do so, the reverse complement sequence of the query file is computed, and `mummer` searches matches between the two forward sequences, as well as matches between the forward sequence of the reference and the reverse complement sequence of the query.
- `-l 20` to filter out matches smaller than 20 bp.

`mummer`'s output consists in a file with three columns representing for each match its position in the query sequence, its position in the reference sequence and its length. The number of matches expected for a random iid sequence grows quadratically with L. For instance one expects $3.5\,10^{18}$ matches of length 2 in a comparison of two sequences of length $L = 10^9$ bp (see Eq. (1)). This explains why one has to define a threshold for the length of matches that `mummer` should output especially when comparing entire eukaryotic genomes. Depending on the length of the sequences to compare, we might vary the value of this threshold.

Logarithmic Binning — Power-laws appear in the tail of distributions, meaning that they are associated to rare events, which are thus subject to strong fluctuations. The high impact of noise in the tail of the distribution can make the assessment of the exponent of the distribution difficult. A way to resolve this issue is to increase the size of the bins with the value of the horizontal axes and renormalise the data accordingly. Namely, the observed values for each bin are divided by the size of the bin. The most common choice to do this is known as the logarithmic binning procedure, which consists in increasing the size of the bin by a constant factor. Note that by doing so, one dramatically reduces the number of data points and some information is lost as one aggregates different data points together in the same bin (and data with different values are summarised together as one data point). Therefore it is often useful to consider both representations, with and without the logarithmic binning; for instance to observe the overrepresentation of matches with length $3n + 2$ for integer n no logarithmic binning should be used. See reference [18] for more details on the logarithmic binning procedure, and on power-law distributions.

Simulating the Evolution of DNA Sequences — To simulate our evolutionary models, we proceeded as follows. A sequence of nucleotides $\mathcal{S} = (s_1, \ldots, s_L)$ of length L with $s_i \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ is evolved through time in small time intervals Δt. The time intervals Δt are small enough such that for all considered evolutionary processes E of our model, which are assumed to occur with

rate ρ_E, we have $\rho_E L \Delta t \ll 1$. At each step, random numbers u_i^E for all positions i and possible evolutionary processes E are drawn from a uniform distribution. The event E then occurs at position i if $u_i^E < \rho_E \Delta t$. These steps are repeated until the desired time t has elapsed.

Sequences evolve according to two simple processes, point mutations and segmental duplications. Point mutations exchange one nucleotide by another and occur with rate μ per bp and unit of time. Note that to mimic the existence of regions under different degrees of selective pressure we allow for regional differences of the point mutation rates. Segmental duplications copy a contiguous segment of K nucleotides starting at position c and paste them to a different position v, such that the K nucleotides at positions v to $v+K-1$ are replaced by the ones from position c to $c+K-1$. As a consequence, the total length of the sequence L stays constant in time. The segmental duplication process occurs with rate λ per bp and per unit of time.

The evolutionary scenario of our simulation has two stages, as shown in Fig. 2. At time $t = 0$, we start with a random iid sequence $\mathcal{S}$ with equal proportions of all 4 nucleotides. During a time interval of length t_0, this sequence evolves according to the two described processes. In this first stage, the mutation rate is the same for all positions. At the end of this stage, the sequence represents the common ancestor of the two species.

At the beginning of the second stage, we duplicate the entire sequence of the common ancestor to generate the genomes of the two species A and B. These sequences are then subdivided into M continuous regions of equal length, in which the point mutation rates $\mu_{A,j}$ (resp. $\mu_{B,j}$) are the same for all sites in a region j and are drawn from the same exponential distribution of mean μ, i.e. the point mutation rate during the first stage. The exponential distribution stipulates the least information under the given constraints.

For simplicity, the length of the M continuous regions is set to K and the segmental duplication rates in both species λ during the second stage are set to zero. Both species then evolve independently for a divergence time t_{sp}, and we compute the MLD from a comparison of the sequences of the two species A and B. Note that even when we chose finite duplication rates after the split (i.e. $\lambda > 0$ in the second stage), the MLDs obtained from the simulated sequences were not qualitatively different.

To control for the potential impact of our choice to keep the genome size constant on our results, we also simulated the evolution of sequences where duplicated segments where added to the sequences (thus generating growing genomes). In that case, duplicates were added at the very end of the sequence, such that duplicates do not disrupt pre-existing matches. This control experiment yields qualitative similar results, in agreement with our theoretical considerations (data not shown).

Simulating the Evolution of DNA Sequences — To simulate our evolutionary models, we proceeded as follows. A sequence of nucleotides $\mathcal{S} = (s_1, \ldots, s_L)$ of length L with $s_i \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ is evolved through time in small time intervals Δt. The time intervals Δt are small enough such that for all considered evolutionary processes E of our model, which are assumed to occur with rate ρ_E, we have $\rho_E L \Delta t \ll 1$. At each step, random numbers u_i^E for all positions i and possible evolutionary processes E are drawn from a uniform distribution. The event E then occurs at position i if $u_i^E < \rho_E \Delta t$. These steps are repeated until the desired time t has elapsed.

Sequences evolve according to two simple processes, point mutations and segmental duplications. Point mutations exchange one nucleotide by another and occur with rate μ per bp and unit of time. Note that to mimic the existence of regions under different degrees of selective pressure we allow for regional differences of the point mutation rates. Segmental duplications copy a contiguous segment

of K nucleotides starting at position c and paste them to a different position v, such that the K nucleotides at positions v to $v+K-1$ are replaced by the ones from position c to $c+K-1$. As a consequence, the total length of the sequence L stays constant in time. The segmental duplication process occurs with rate λ per bp and per unit of time.

The evolutionary scenario of our simulation has two stages, as shown in Fig. 2. At time $t=0$, we start with a random iid sequence $\mathcal{S}$ with equal proportions of all 4 nucleotides. During a time interval of length t_0, this sequence evolves according to the two described processes. In this first stage, the mutation rate is the same for all positions. At the end of this stage, the sequence represents the common ancestor of the two species.

At the beginning of the second stage, we duplicate the entire sequence of the common ancestor to generate the genomes of the two species A and B. These sequences are then subdivided into M continuous regions of equal length, in which the point mutation rates $\mu_{A,j}$ (resp. $\mu_{B,j}$) are the same for all sites in a region j and are drawn from the same exponential distribution of mean μ, i.e. the point mutation rate during the first stage. The exponential distribution stipulates the least information under the given constraints.

For simplicity, the length of the M continuous regions is set to K and the segmental duplication rates in both species λ during the second stage are set to zero. Both species then evolve independently for a divergence time t_{sp}, and we compute the MLD from a comparison of the sequences of the two species A and B. Note that even when we chose finite duplication rates after the split (i.e. $\lambda > 0$ in the second stage), the MLDs obtained from the simulated sequences were not qualitatively different.

To control for the potential impact of our choice to keep the genome size constant on our results, we also simulated the evolution of sequences where duplicated segments where added to the sequences (thus generating growing genomes). In that case, duplicates were added at the very end of the sequence, such that duplicates do not disrupt pre-existing matches. This control experiment yields qualitative similar results, in agreement with our theoretical considerations (data not shown).

Estimating the value of the Power-law Exponent

To estimate the value of the exponent of the power-law α, we compute the maximum likelihood estimator. The estimator $\hat{\alpha}$ is simply the value of α that maximizes the log likelihood $\mathcal{L}$:

$$\mathcal{L} = \sum_{i=1}^{n} \left[\ln(\alpha - 1) - \ln(x_{min}) - \alpha \ln\left(\frac{x_i}{x_{min}}\right) \right], \tag{12}$$

such that

$$\hat{\alpha} = -1 - n \left[\sum_{i=1}^{n} \ln\left(\frac{x_i}{x_{\min}}\right) \right]^{-1}, \tag{13}$$

while the value of $x_{\min}$ has to be determined visually. This estimator is also sometimes referred to as the Hill estimator [12].

To estimate the robustness of the value of the exponnents found using this method, we proceeded to block boostrap experiments on the Human Mouse exome comparison. For each boostrap, we sampled 5% of the mouse exons, and compared them to all human exons. In each experiment, we calculated the exponnent of the MLD using the maximum likelihood estimator as described in [18]. We repeated this procedure a 100 times. Values of α were all in the range $[-4.7, -5.2]$ and the mean value for the exponent was $\alpha = -4.9$.

References

[1] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–410.

[2] Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., and Haussler, D. (2004). Ultraconserved elements in the human genome. *Science*, 304(5675):1321–1325.

[3] Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2015). Ensembl 2015. *Nucleic Acids Research*, 43(D1):D662–D669.

[4] Dalquen, D. A., Altenhoff, A. M., Gonnet, G. H., and Dessimoz, C. (2013). The impact of gene duplication, insertion, deletion, lateral gene transfer and sequencing error on orthology inference: a simulation study. *PloS one*, 8(2):e56925.

[5] Dalquen, D. A., Anisimova, M., Gonnet, G. H., and Dessimoz, C. (2012). Alf — a simulation framework for genome evolution. *Molecular biology and evolution*, 29(4):1115–1123.

[6] Fitch, W. M. (2000). Homology: a personal view on some of the problems. *Trends in genetics*, 16(5):227–231.

[7] Flanders, H. (1973). Differentiation under the integral sign. *The American Mathematical Monthly*, 80(6):pp. 615–627.

[8] Gabaldón, T. and Koonin, E. V. (2013). Functional and evolutionary implications of gene orthology. *Nature Reviews Genetics*, 14(5):360–366.

[9] Gao, K. and Miller, J. (2011). Algebraic distribution of segmental duplication lengths in whole-genome sequence self-alignments. *PLoS ONE*, 6(7):e18464.

[10] Gao, K. and Miller, J. (2014). Human–chimpanzee alignment: Ortholog exponentials and paralog power laws. *Comput Biol Chem*, 53:59–70.

[11] Han, M. V., Demuth, J. P., McGrath, C. L., Casola, C., and Hahn, M. W. (2009). Adaptive evolution of young gene duplicates in mammals. *Genome research*, 19(5):859–867.

[12] Hill, B. M. et al. (1975). A simple general approach to inference about the tail of a distribution. *The annals of statistics*, 3(5):1163–1174.

[13] Kuhn, W. (1930). Über die Kinetik des Abbaues hochmolekularer Ketten. *Berichte der Deutschen Chemischen Gesellschaft*, 63:1502–1509.

[14] Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biology*, 5(2):R12.

[15] Massip, F. and Arndt, P. F. (2013). Neutral evolution of duplicated dna: an evolutionary stick-breaking process causes scale-invariant behavior. *Physical Review Letters*, 110(14):148101.

[16] Massip, F., Sheinman, M., Schbath, S., and Arndt, P. F. (2015). How evolution of genomes is reflected in exact DNA sequence match statistics. *Molecular Biology and Evolution*, 32(2):524–535.

[17] Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3):443–453.

[18] Newman, M. E. (2005). Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46(5):323–351.

[19] Panchin, A. Y., Gelfand, M. S., Ramensky, V. E., and Artamonova, I. I. (2010). Asymmetric and non-uniform evolution of recently duplicated human genes. *Biology direct*, 5(1):54.

[20] Pegueroles, C., Laurie, S., and Albà, M. M. (2013). Accelerated evolution after gene duplication: a time-dependent process affecting just one copy. *Molecular biology and evolution*.

[21] Salerno, W., Havlak, P., and Miller, J. (2006). Scale-invariant structure of strongly conserved sequence in genomic intersections and alignments. *Proceedings of the National Academy of Sciences*, 103(35):13121–13125.

[22] Scannell, D. R. and Wolfe, K. H. (2008). A burst of protein sequence evolution and a prolonged period

of asymmetric evolution follow gene duplication in yeast. *Genome research*, 18(1):137–147.

[23] Sheinman, M., Massip, F., and Arndt, P. F. (2015). Statistical properties of pairwise distances between leaves on a random yule tree. *PLoS ONE*, 10(3):e0120206.

[24] Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197.

[25] Studer, R. A. and Robinson-Rechavi, M. (2009). How confident can we be that orthologs are similar, but paralogs differ? *Trends in Genetics*, 25(5):210–216.

[26] Ziff, R. M. and McGrady, E. D. (1985). The kinetics of cluster fragmentation and depolymerisation. *Journal of Physics A: Mathematical and General*, 18:3027.