Universality in Large-Scale Structure of Complete Genomes

Li-Ching Hsieh*‖, Chang-Heng Chang*, Ta-Yuan Chen*, Wen-Lang Fan* and H.C. Lee*†‖
*Department of Physics, †Department of Life Sciences and ‖Center for Complex Systems,
National Central University, Chungli, Taiwan 320
hclee@phy.ncu.edu.tw http://sansan.phy.ncu.edu.tw/∼hclee/

Abstract: - The frequency of distributions of oligonucleotides, or k-spectra, of DNA sequences are often much wider than those expected of compositionally similar random sequences. Here we made a systematic study of the widths of k-spectra of oligonucleotides two to ten nucleotides long from all extant prokaryotic and eukaryotic complete genomes and found that, with the exception of one organism, the k-spectra of all the complete genomes belong to one universality class that is expressed in a an exceedingly simple formula. The genomes are far from being random sequences, yet they exhibit strong self-similarity and intervals among identical oligonucleotides conforms to expectation for uncorrelated events. We show that a simple genome growth model using stochastic segmental duplication as a major mode of growth can reproduce the genomic data - including the k-spectra, the universality classes and the self-similarity - in detail. We discuss possible connections between our findings and the notion that early genomes developed in a proteinless RNA world.

Key-Words: - Genome analysis, statistical properties, evolution, RNA world, genome growth model

Genomes have been analyzed as texts of four nucleotides - A, C, G and T - to look for statistical patterns [1, 2], to construct a dictionary [3, 4] and to study phylogeny and evolution [5, 6], among others. Such an analysis may begin by counting the number of times every oligonucleotide, or word, of a given length (k-mers for k-letter words) occur in the genome. If the occurrence frequency is considered as light frequency and the number of k-mers with that occurrence frequency as light intensity, *i.e.*, number of photons, then a distribution of occurrence frequency can be considered analogously to a standard optical spectrum. We henceforth call such a distribution a k-spectrum. The 1-spectrum of a genome gives the four numbers forming its base composition. The k-spectra of a genome for all k's may say something about the genome's large-scale structure. The k-spectra of many genomes for a single k may reveal properties shared by genomes. Here we report the result of a systematic study of a single k-spectrum property, namely its generalized spectral width (see below) relative to that of a corresponding random sequence. The study's scope covers all complete prokaryotic genomes (as of April 2003) and all chromosomes from complete eukaryotic genomes (July 2003) [7], for k=2 to 10. Our findings are unanticipated. All data (from all sequences and all k's) except those from the genome of the malaria causing parasite *Plasmodium falciparum* form a universality class described by a two-parameter exponential equation with k being the exponent. Data from the 14 chromosomes of *Plasmodium* form a related but distinct class.

Fig. 1 shows three sets of 5-spectra and gives a general overview of k-spectra for short words. The plots in black in Fig. 1 are the 5-spectra of three representative complete genomes with different percentages of A+T content, or p, and the plots in green show the 5-spectra of corresponding random sequences obtained by thoroughly scrambling the genomes (the spectra in orange are from sequences generated in a growth model, see below). Depending on the value of p, the random spectra are composed of one or more very narrow subspectra while each of the genomic spectra essentially comprises a single, much wider distribution. In the general case, the k-spectrum of a sequence that has approximately the same number of A and T bases and approximately the same number of G and C bases - true for most genomes - is composed of k+1 subspectra that (i) are narrow if the sequence is random and broad and overlapping if it is genomic, and (ii) coalesce into a unimodal spectrum as p approaches 0.5 (see supplementary information). In this regard the (B) and (C) panels in Fig. 1 may both be viewed qualitatively as superpositions of six sets of shifted (A)'s.

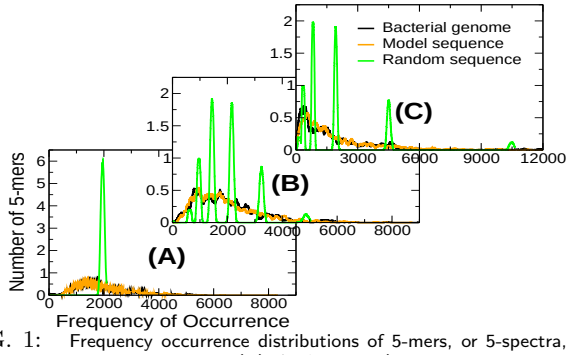

FIG. 1: Frequency occurrence distributions of 5-mers, or 5-spectra, from three representative prokaryotes. (A) *A. fulgidus* (with A+T content p=0.5); (B) *S. pneumoniae* (p=0.6); (C) *C. acetobutylicum* (p=0.7). Abscissa give occurrence frequency and ordinates give number of 5-mers. The black, green and orange curves are respectively from the complete genomes, the randomized genome sequences and sequences generated in a model described in the text. The spectra shown have been forward and backward averaged to reduce fluctuation.

We quantify the broadening of a genomic k-spectrum relative to its random-sequence counterpart as follows. For each k-spectrum of a complete genome a *reduced spectral width*, $\mathcal{M}_\sigma$, defined as the weighted average of $(\sigma_m/\sigma'_m)^2$, is computed, where σ'_m is the expected width of the m^{th} subspectrum of the random sequence and σ_m is the width of the corresponding (underlying) subspectrum of the genome sequence. The weight for each subspectrum is given by the number of k-mers in the subspectrum. In the simplest case, when the k-spectrum of the random sequence is unimodal ((A) in Fig. 1) - occurring when p is close to 0.5 - $\mathcal{M}_\sigma$ is just the square of the ratio of the widths of the genomic and random k-spectra. The quantity $\mathcal{M}_\sigma$ has some interesting properties that will become clear later.

The reduced spectral widths of the k-spectrum, k=2 to 10, of 108 complete prokaryotic genomes (the prokaryotes) and 113 complete chromosomes from ten eukaryotic genomes (the eukaryotes) are shown against sequence

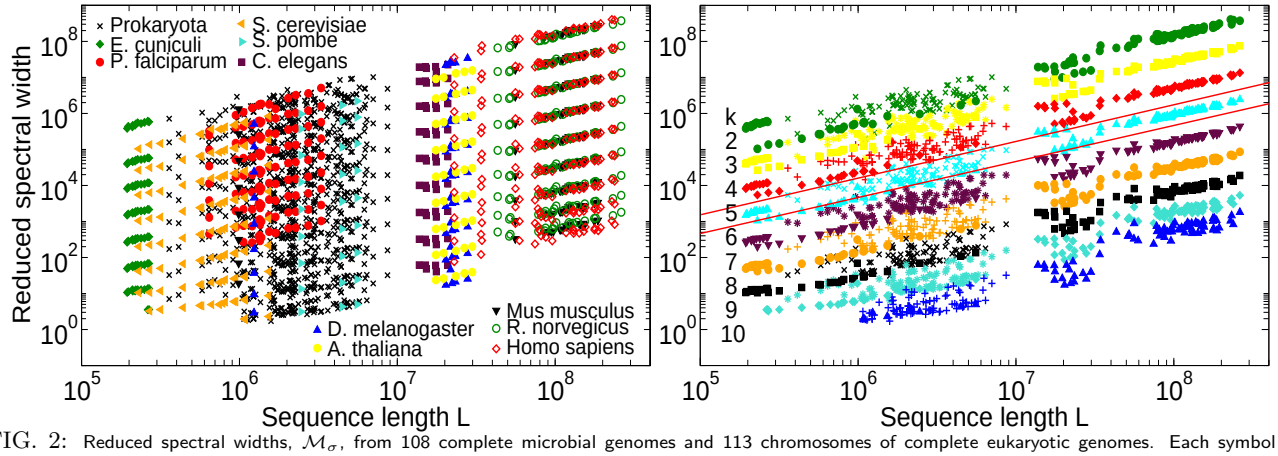

FIG. 2: Reduced spectral widths, $\mathcal{M}_\sigma$, from 108 complete microbial genomes and 113 chromosomes of complete eukaryotic genomes. Each symbol is the $\mathcal{M}_\sigma$ value of one k-spectrum from one complete sequence. Left panel, $\mathcal{M}_\sigma$ color-coded by organism; right panel, $\mathcal{M}_\sigma$ color-coded by k, excluding data from 14 chromosomes of *P. falciparum*, where each "k-band" contains data from 207 complete sequences. Data have been multiplied by factor of 2^{10-k} to delineate the k-bands for better viewing. Data for which $4^k > L$, when $\mathcal{M}_\sigma \approx 1$ regardless of sequence content, have been discarded. Straight red lines in the plots are $\mathcal{M}_\sigma \propto L$ lines.

length (L) as log-log plots in the two panels of Fig. 2. Each datum gives the $\mathcal{M}_\sigma$ value of one k-spectrum from a sequence. In the left panel the data are color-coded by organisms. Data from the three mammals, human (orange ◇), mouse (▼) and rat (green ◇) are practically superimposed, showing that the present analysis is insensitive to whatever mutations, from large chromosomal segment exchanges to gene-modifying point mutations, that may have caused closely related organisms to diverge. One general trend of the results is immediately evident: the data form bands running linearly with L.

In the right panel where data from *Plasmodium* are excluded, the data are color-coded after k. Here a second prominent feature is that all data from the 207 sequences for a given k form a *k-band* that runs along a straight $\mathcal{M}_\sigma \propto L$ line despite heterogeneity in length (0.2 Mb to 0.3 Bb) and base composition ($p_{AT} = 0.2$ to 0.8) of the sequences. For instance, data from the mammalian chromosomes share the same line with data from the single-celled parasite *E. cuniculi* whose chromosomes are a thousand times shorter. A third feature is that the bands are close to being equally spaced, with the lower k-bands lying higher. The average $\mathcal{M}_\sigma$ value for the 2-band is about 2000 times that of the 10-band.

The complete genomes are self-similar. Fig. 3 (A) shows two arbitrarily selected examples of this phenomenon, respectively occurring in the 0.29 Gb *Rat Chromosome I* (+) and in the 4.6 Mb *E. coli* genome (•). In each case, sets of eight segments of lengths equal to $1/2^n$ the full length, n= 2 to 6 for *E. coli* and to 10 for *Rat Chr. I*, are randomly spliced from the full genome/chromosome and their $\mathcal{M}_\sigma$ are computed and plotted against segment length. In the plots, the full-length set include the original genome/chromosome and randomly selected segments, one from each of the segment sets that are not shorter than 1 Mb, replicated to full length. Virtually all the segments have the same $\mathcal{M}_\sigma$ to L ratio as the parent sequence. In the case of *E. coli* small deviations are observed only when segment lengths are shorter than 4^k. These results suggest that complete genomes/chromosomes emulate self-organized critical systems [8].

Although the complete genomes are far removed from being random sequences, k-mers are distributed along a complete genome essentially as uncorrelated objects. The interval distributions for three 4-mers with respective mean intervals approximately equal, twice and half the overall mean interval for all 4-mers in *E. coli* but otherwise randomly selected are shown in Fig. 3 (B). Each distribution is a simple exponential - with an exponent inversely proportional to the mean interval - typical of uncorrelated events. Thus genomes show clear signs of having been stochastically generated.

The linearity of the k-bands in Fig. 2 implies that for given k, the quantity $L_r(k) = \mathcal{M}_\sigma / L$, or *effective root-sequence length*, is an approximate universal constant. In Fig. 3 (C) the black symbols give values for $L_r(k)$ averaged over k-bands. The geometric recursion formula

$$L_r(k)/L_r(k-1) = t; \quad 3 \le k \le 10 \qquad (1)$$

summarizes the combined prokaryote (▲ in Fig. 3 (C)) and eukaryote (*Plasmodium* excluded; ■) data, where $L_r(2)$=300±180 b and t=2.58±0.44, and reduces the 1893 pieces of data in Fig. 2 (right panel) to two universal constants. We refer to the data represented by Eq. (1), whose mean is the straight line in Fig. 3 (C), as a universality class. Data for the fourteen chromosomes of *Plasmodium* (• in Fig. 3 (C)), the sole exceptions to this class, form a related but distinct class given by $L_r(2)$=270±120 b and t=1.38±0.04. The two classes differ only in the value of t, and we refer to the value $L_r(2) \approx 300$ b common to both classes as the universal root-sequence length.

The 108 prokaryotes are relatively homogeneous in length - 0.4 to 7 Mb - but highly heterogeneous in percentage AT content, or $p(AT)$ - 26% to 0.75%. The inverse is the case for the eukaryotes where length ranges from 0.2 Mb to 300 Mb and $p(AT)$ ranges from 53% to 64%. The exception is *Plasmodium* whose $p(AT)$ is 81±1% [9]. This alone does not explain why *Plasmodium* should form its own universality class because some prokaryotes in the dominant class also have extremely biased base compositions - *U. urealyticum* and *B. aphidicola* are 75% AT and *S. coelicolor* is 72% CG.

About 85% of a prokaryote is comprised of coding regions, whereas on the whole most of an eukaryotic chro-

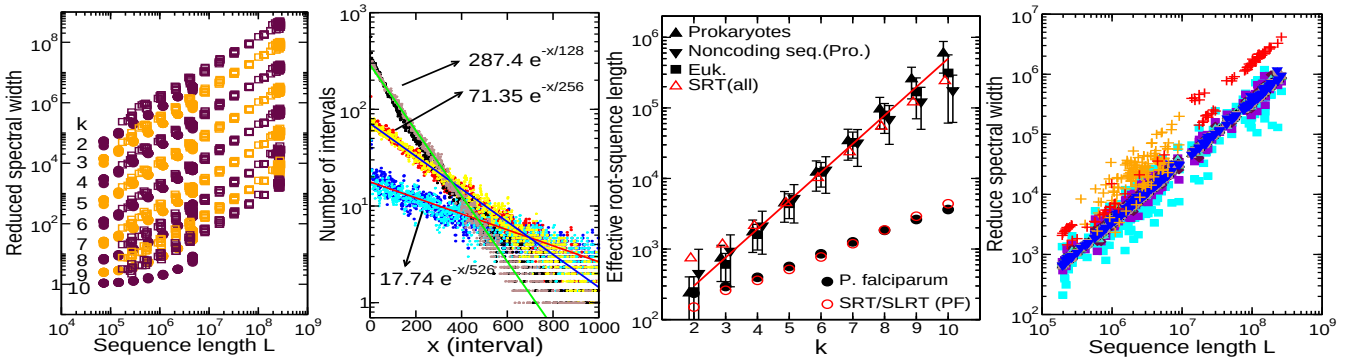

FIG. 3: (A) Self-similarity of complete genomes. $\mathcal{M}_\sigma$ color-coded by k of randomly selected segments 2^n-*th* the full length of genomes of *Rat Chromosome I* (+'s) and *E. coli* (•'s); (B) Interval distributions of three randomly selected 4-mers in the genome (dark color) and model sequence (light color) for *E. coli* with respective mean intervals approximately equal (red/gold), twice (black/gray) and half (navy/sky) the overall mean interval for all 4-mers in *E. coli*. The straight lines are the (exponential) distributions expected of uncorrelated events. (C) Effective root-sequence lengths L_r. Each piece of data (with error flags) is obtained from averaging $L/\mathcal{M}_\sigma$ over a k-band such as those seen in Fig. 2. Black symbols are from genomic data: ▲, prokaryotes; ▼, non-coding regions in prokaryotes; ■ eukaryotes; • *Plasmodium*; The red line gives the mean of the relation Eq. (1). Orange symbols show results obtained from model sequences: △, RPM model for prokaryotes (L_0=30, R=75); ▲, SRT model for prokaryotes (L_0=200, $\bar{l}$=20±12); ■, SRT/SLRT model for eukaryotes ($\bar{l}_L$=5000±5000); •, SRT/SLRT model for Plasmodium (same as above but $\bar{l}$=60±36 in SRT phase). (D) Reduced spectral widths for k-spectra, k=2 to 10, of 207 sequences in the control set (dark colored symbols) and for 2-spectra (multiplied by factor 3) of complete prokaryotes (yellow) and eukaryotes (green).

mosome is non-coding (coding regions make up less than 2% of the human genome). In Fig. 3 (C) the $L_r(k)$ for sequences obtained by concatenating the noncoding segments in prokaryotes are shown as ▼. Compared to the their overall k-dependence, the difference between the data for whole genomes and for noncoding segments is secondary and one may infer that no essential difference between coding and noncoding regions obtains. (Codons must be read in the open reading frame - one of six possible frames - and they seem not to make a major impact on 3-spectra.)

In what follows we say an "m-replica" is generated by replicating m times a random "root"-sequence. The reduced spectral width $\mathcal{M}_\sigma$ was defined such that its expected value for any k-spectrum of an m-replica is m. We have generated a control set of 207 sequences whose members are replicas with 300 b random root-sequences such that, corresponding to a genomic sequence of length L and base composition p in the combined set of prokaryotes and eukaryotes (*Plasmodium* excluded), there is a $L/300$-replica with the base compositions in the control set. The $\mathcal{M}_\sigma$ of the k-spectra, k=2 to 10, for all sequences in the control set were computed and the results shown in Fig. 3 (D) (symbols in dark color) confirm expectation: $\mathcal{M}_\sigma$ approximately equals $L/300$, respectively, independent of p and k. Since it is very difficult to greatly increase the value of $\mathcal{M}_\sigma$ by any means other than replication/duplication, we infer the following from these results: a large value for $\mathcal{M}_\sigma$ suggests a correspondingly large amount of replication in the sequence and, for a set of sequences, linear dependence of $\mathcal{M}_\sigma$ on sequence length suggests a common root-sequence length for the set. If it is assumed that a sequence having a large $\mathcal{M}_\sigma$ is a result of replication/duplication from a random root-sequence, then the root-sequence must not be longer than $L/\mathcal{M}_\sigma$, because any randomizing action on a sequence reduces $\mathcal{M}_\sigma$. Hence our usage of the term effective root-sequence length for $L_r(k)$. Because the data show $L_r(k)$ increasing with k, the shortest root-sequence length, $L_r(2)$, is the crucial length. From the data and the universality class we can infer that a very high degree of replication/duplication occurred during the formation of the complete genomes under study and that this process began when the genomes' ancestors were not longer than the universal root-sequence length of 300 b.

In Fig. 3 (D) the data for the control set closely follow the k=2 band of genomic data, shown again as yellow (prokaryotes) and green (eukaryotes) symbols (here for better visibility: boosted by a factor of 3). However, the presence of a strong k-dependence in the genome data (right panel, Fig. 2) and its absence in the control set data rule out the possibility that the complete genomes/chromosomes are simple replicas, a finding independently corroborated by other sources [10, 11]. The observed k-dependence need be accounted for by another mechanism of genome growth. We have inquired into two mechanisms that entail the elements of duplication and stochasticity in differing combinations. Each has been incorporated into a simple two-parameter growth model: (i) Replication plus random mutation (RPM) model where a random sequence (with preselected p) of length L_0 is replicated to full genome length, followed by random point mutations at the rate of R mutations per 100 b sequence length; (ii) Stochastic replicative transposition (SRT) model in which an initial random sequence of length L_0 is grown via repeated duplication of randomly selected short segments of average length $\bar{l}$ and random transposition. The models generate random sequences in the large R and small $\bar{l}$ limits and simple replicas in the opposite limits. In an extension of SRT, called SLRT, a sequence first generated in SRT to a length of 1 Mb is further extended by stochastic replicative transposition, but with long segments whose average $\bar{l}_L$ is much greater than $\bar{l}$.

Some of the modeling efforts are shown as orange symbols in Fig. 3 (C). The data (black symbols) in Fig. 3 (C) impose severe constraints on the parameters of the model. The initial length L_0 needs to be a less than $L_r(2)\approx$300 b and the slope of the logL_r versus k curve puts strict limits on the mutation rate R and average length of duplicated segments $\bar{l}$ - a larger slope requires a larger R and/or a smaller $\bar{l}$. Curiously, the value of $\bar{l}$ (20±12 b) is very close to the size (18-25 nt) of microRNAs, small regulatory RNAs that have recently been found to be prolific at least in eukaryotes [12]. Given their extreme simplicity, the model results reproduce the

main features of the genome data reasonably well. The SRT results are slightly better than those of RMP. The SRT and SRT/SLRT model sequences reproduce the k-spectra (orange in Fig. 1), self-similarity (Fig. 3 (A)) and interval distributions (Fig. 3 (B)) of the genomes. The simplicity of the growth mechanisms and the fact that they are driven by stochastic events underlie the robustness of the results and may explain the emergence of the universality classes. Their heuristic utility notwithstanding, neither RPM nor SRT in isolation provides a realistic model of the early large-scale growth of genomes; a realistic model should be closer to being a hybrid. It is to be understood that in any such model all the key elements - genome replication, point mutation, replicative transposition - are mutation events and are subject to the usual rule of natural selection: only non-deleterious events can be fixed.

There is abundant evidence that genomes used large duplications as one mode for its growth: the large amounts of repeats in both prokaryotes [13] and eukaryotes [14, 15]; the preponderance of homologous genes and pseudogenes in all life forms [16]; chromosome segment exchanges that to seem characterize mammalian [17] and plant [18] radiations. Not surprisingly such a growth strategy has been shown to have the ability to greatly enhance the rate of evolution [19] and to increase the robustness of organisms [20]. Our study suggests that duplication may have been a dominant mode of growth very early on in the life of genomes, mostly likely when genomes were not more than 300 b long. This in turn may be taken as at least weakly supporting the conjecture that the ancestors of present-day genomes started life in the proteinless RNA world [21]. A genome not more than 300 nt in length would have been large enough to harbor a sufficient number of small ribozymes (of the order of 30 nt [22]) providing it the ability to replicate reliably but would have been too small to encode enzymatic proteins. The SRT/SLRT model may be understood as a coarse-grain representation of genome growth in a world in which primitive genomes evolved, diverged and grew via small duplications (presumably with the aid of small ribozymes), and later (with the arrival of proteins and enzymes) acquired the ability to undergo duplication on much larger scales. How realistic is such a picture and how such duplications were modified and fine-tuned by point mutations to enhance species fitness and affect species diversion are issues still to be settled.

Acknowledgment. This work is partly supported by the National Science Council (ROC) grant NSC 92-2119-M-008-012 to HCL.

[1] Karlin S. *et al.* Statistical analyses of counts and distributions of restriction sites in DNA sequences. *Nucl. Acids Res.* **20**, 1363-1370 (1992).

[2] Smith H.O. *et al.* Frequency and distribution of DNA uptake signal sequences in the *Haemophilus influenzae* Rd genome. *Science* **269**, 538-540 (1995).

[3] van Helden J., Andre B. & Collado-Vides J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.* **281**, 827-842 (1998).

[4] Bussemaker H.J., Li H. & Siggia E.D. Building A Dictionary for Genomes: Identification of Presumptive Regulatory Sites by Statistical Analysis. *PNAS* **97**, 10096-10100 (2000).

[5] Karlin S. & Burge C. Dinucleotide relative abundance extremes: a genomic signature. *Trends in Genetics* **11** 283-290 (1995).

[6] Qian J., Luscombe N.M., Gerstein M. Protein family and fold occurrence in genomes: power-law behavior and evolutionary. *J. Mol. Biol.* **313**, 673-681 (2001).

[7] All complete sequences are take from GenBank. prokaryotes: www.ncbi.nlm.nih.gov/genomes/Complete.html (April 2003); Eukaryotes: .../genomes/static/euk_g.html (July 2003)

[8] Bak P., Tang C. & Wiesenfeld K. Self-Organized Criticality - An explanation for 1/f noise. *Phys. Rev. Lett.* **59**, 381-384 (1987).

[9] Gardner M.J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498-511 (2002).

[10] Ohno S, Evolution by gene duplication. (Springer, New York, 1970).

[11] Hughes A.L., da Silva J. & Friedman R. Ancient genome duplications did not structure the Human Hox-bearing chromosomes. *Genome Res.* **11**, 771-780 (2001).

[12] Ambros, V. MicroRNA pathways in flies and worms: growth, death, fat, stress, and timing. *Cell* **113**, 673-676 (2003).

[13] Lars Juhl Jensen, L.J. *et al.* Three views of microbial genomes. *Res. Microbiol.* **150**, 773-777 (1999).

[14] Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

[15] Venter, J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304-1351 (2001).

[16] Meyer A. Duplication, duplication. *Nature* **421**, 31-32 (2003).

[17] O'Brien S.J. *et al.* The Promise of Comparative Genomics in Mammals. *Science* **286**, 458-481 (1999).

[18] Grant D. *et al.* Genome organization in dicots: Genome duplication in Arabidopsis and synteny between soybean and Arabidopsis. *PNAS* **97**, 4168-4173 (2000).

[19] Zhang, Y-X. *et al.* Genome shuffling leads to rapid phenotypic improvement in bacteria. *Nature* **415**, 644-646 (2002).

[20] Gu, Z. *et al.* Role of duplicate genes in genetic robustness against null mutations. *Nature* **421**, 63-66 (2003).

[21] Joyce G.F. The antiquity of RNA-based evolution. *Nature* **418**, 214-221 (2002).

[22] Forster A.C. & Symons R.H. Self-cleavage of plus and minus RNAs of a virusoid and a structural model for the active sites. *Cell* **49**, 211-220 (1987).