BIOINFORMATICS

Vol. 00 no. 00 2007
Pages 1–7

Numerical models for genome growth

Jan Wigger[a], Sing-Guan Kong, Hong-Da Chen[b], Wen-Lang Fan[b], H.C. Lee[b], Andrew E. Torda[a*]

[a]University of Hamburg, Centre for Bioinformatics, Bundesstrasse 43, 20146 Hamburg, Germany
[b]Graduate Institute of Systems Biology and Bioinformatics, National Central University, Chungli, Taiwan 320

ABSTRACT

Motivation: Modern genomes and their statistical properties reflect the series of steps by which they evolved. A very simple model for genome growth has been implemented, compared to other literature models and properties of bacterial genomes. The model has four types of nucleotide and adjustable rates of mutation, deletion and insertion of single bases and triplets, as well as a simple mechanism for segmental duplication.

Results: The amount of order/structure/correlation within a genomic sequence has often been attributed to a balance of duplication and mutations. This model supports this, but we find that many results are explained equally well by the presence of a bias in the selection of codons. Other models have built correlations into the generation of DNA sequences. This appears to be unnecessary when using simple, but biologically plausible evolutionary mechanisms.

Availability: The source code for the simulations is available from www.uni.uni-hamburg.de/torda/genome_sim

Contact: torda@zbh.uni-hamburg.de

1 INTRODUCTION

If genomes were random sequences, their statistical properties could be predicted from probabilistic arguments. Instead, genomes have patterns which may be interpreted in terms of short- or long-range correlations, fluctuations, information content or noise properties (Chen *et al*, 2004, 2005a; Chen *et al*, 2005b; Li *et al*, 1994, 1995; Messer *et al*, 2005a,b; Peng *et al*, 1992; Voss , 1992). The aim of this work has been to use brute force simulations to show the effects of some evolutionary mechanisms on the statistical properties of genomic sequences.

The genomes we see today reflect biochemical mechanisms of change (deletions, duplication, mutations), but filtered by selection pressures and fixation probabilities. For example, single base insertions or deletions may occur frequently, but if they cause frame shifts, their contribution to modern genes will be small. Triplet changes (codons) may occur much less often, but within genes, they may be tolerated and become fixed in a population. In this work, we only model accepted events. We do not separate mechanisms from evolutionary pressure. The model should reproduce gross statistical properties, but it does not touch on the underlying biochemistry or population properties.

One can then define the simulation framework and some limitations. Our genome is a string of bases with regions labelled as genes or non-genes. Single bases or triplets may be inserted, deleted or mutated with adjustable probabilities. Single bases may not be inserted or removed from a region labelled as a gene. Duplications occur by copying a segment and inserting it into a separate region in the genome.

This approach has as its basis earlier literature simulations. Messer *et al* (2005a,b) employed a binary alphabet (two types of base) and adjustable rates for single base mutations, insertions and deletions as well as duplication of bases. Their model had the interesting restriction that duplicated bases were placed adjacent to the source. This has obvious implications when assessing results in terms of short- and long-range correlations. This rather artificial restriction did, however, allow for an elegant analytical interpretation of their numerical results.

Several authors have taken a more abstract approach, observed long-range correlations and then built a sequence generator which incorporates the appropriate probabilities (de Souza and Anteneodo, 1995; Tai *et al*, 2006). This could be seen as similar to a low-order Markov model or a sequence with a rapid decay of a memory function. It could be labelled a model for genome growth. Others have demonstrated that genome-like qualities can be put in a sequence by Fourier transforming a sequence, manipulating in reciprocal space and back transforming (Buldyrev *et al*, 1995). This has been seen as the generation of control sequences. Most of this earlier work has used just two base types as it is adequate for demonstrating certain patterns in genomes.

In this work, a very simple model was used, although it had 4 types of base. A set of evolutionary mechanisms was then coded. Firstly, segments could be labelled as coding or non-coding. This should not be interpreted too literally. It simply means that single base insertions or deletions were forbidden in regions labelled as coding. Secondly, a possible source of order is brought about by selection pressures at the protein level. Therefore, the model could include a bias in the quasi-random selection of triplets/codons. Since this is an adjustable parameter, the model could be reduced to calculations without this bias. Unlike the statistical models, correlations between adjacent bases were never explicitly created. Each base or triplet was be inserted without checking its preceding or succeeding residues. If correlations appear in the final result, they have not been deliberately constructed.

A goal of this work is to compare against real genomes and there is no limit to the possible measures. For this model, we are interested in properties related to short or long range order, information or entropy. We implemented three literature methods. Firstly, a physicist might want to use a simple auto-correlation function. We tested the method of Messer *et al* (2006). Next, one is interested in long-range order, so a method introduced by Peng *et al* (1992) was used

*to whom correspondence should be addressed

© Oxford University Press 2007.

with detrended fluctuation analysis since this should account for patchiness in sequences (Peng *et al*, 1994). This treats a genome as a walk along a purine/pyrimidine "space" and measures the fluctuations taken by the sequence. In a random sequence, these are not correlated, but if long-range correlations are present, they can be recognized by a distinct power law exponent. Finally, we used an approach with its philosophy based on information theory. Chen *et al* (2004) defined a fragment or word length (typically 4 to 10) and then measured a property which could be labelled entropy or Shannon information if these words are considered states of a system. If we use the word entropy, these authors noted that real genomes seemed to have an amount of "entropy" which scales linearly with genome size and appeared to be a multiple of some minimum information due to a sequence with a length, L_r or minimum root sequence length. However one labels or interprets the quantity, it is important as it seems to be almost independent of organism, but very different between real genomes and simple random sequences. L_r can be calculated by a simple formula given in Methods (Chen *et al*, 2005b; Chen *et al*, 2005a).

2 METHODS

2.1 Model

A genome was represented by a string using the alphabet of the four nucleotide bases. Each position was labelled as coding or non-coding and with a reading direction. Initial sequences were generated by specifying the number of coding n_c and non-coding n_{nc} regions as well as their initial sizes. Coding or non-coding segments were then added according to probabilities $n_c/(n_c+n_{nc})$ or $n_{nc}/(n_c+n_{nc})$ respectively until the desired number was reached. For the generation of coding regions and during triplet insertion or mutation (described below) triplets were chosen quasi-randomly, but with biases taken from a vector $\vec{p}_c$ of codon probabilities with the obvious normalization that the elements of $\vec{p}_c$ sum to 1. Three possibilities were used. Firstly, all elements of $\vec{p}_c$ could be set equal (unbiased). Secondly, the probabilities were taken from values for *Escherichia coli* (Nakamura *et al*, 1998). Thirdly, an artificial, but plausible set of probabilities was generated by taking each element of $\vec{p}_c$, replacing it by the element squared and normalizing so the probabilities again sum to 1. This is referred to as the emphasised *E. coli* bias. This does not model any known distribution, but serves to show the effects of codon bias.

Given an initial sequence, each simulation followed the stochastic method of Gillespie (1977). A set of possible moves is described and each assigned a probability. At each simulation step one of the moves is chosen according to its probability. In these calculations a move was attempted at each step, unlike Gillespie's method which attempts to reproduce a timescale. Compared to the published method of Gillespie, temporal order was preserved, but not temporal scale.

Seven operations on the sequence were defined:

- `Insertion 1`: Insertion of a nucleotide into a non-coding region
- `Insertion 3`: Insertion of a codon into a coding region
- `Deletion 1`: Deletion of a nucleotide from a non-coding region
- `Deletion 3`: Deletion of a codon from a coding region
- `Mutation 1`: Mutation of a nucleotide from a non-coding region
- `Mutation 3`: Mutation of a codon from a coding region
- `Segmental Duplication`: Duplication of a segment with length taken from a Gaussian distribution of specified mean and variance. Insertion into a coding segment was forbidden. Intact coding regions in inserted segments remained coding regions. Partial copies of coding regions were marked as non-coding.

2.2 Statistical Measures

Three measures were used to quantify the statistical properties of sequences.

Algorithm 1 Model algorithm

```
set up initial sequence
while number of steps ≤ maximum number of steps do
    choose 1 of the 7 operations with specified probability (rates)
    execute operation at valid random site
end while
```

Effective root sequence length, L_r The effective root sequence length L_r was calculated according to Chen *et al* (2004) with a correction for uneven purine/pyrimidine ratio (Chen *et al* (2005a)). The sequence was broken into all overlapping words of length k, and the frequency, f, of each word counted.

$$L_r(k) = \frac{b_k 4^k (\overline{f})^2}{\sigma^2} \tag{1}$$

where k is the word length, $\overline{f}$ is the average occurrence of a word with k letters in a sequence, σ is the standard deviation of word frequencies and $b_k = 1 - 1/2^{(k-1)}$

Detrended fluctuation analysis Detrended Fluctuation analysis was proposed as a method to measure long-range power law correlations in sequences by Peng *et al* (1994) and is based on binary (purine/pyrimidine) statistics.

A walker *u* indicates the net difference between purine and pyrimidines in a region of length *l*. The walker starts at position *i* and walks down the sequence for *l* bases. Each time it encounters a purine (A or C) *u(i)* = 1, otherwise, when having found a pyrimidine, *u(i)* is -1.

y(l) is the net displacement of a segment of length *l* and is defined as

$$y(l) = \sum_{i=1}^{l} u(i). \tag{2}$$

The entire sequence of length *L* is divided into $\frac{L}{l}$ non-overlapping segments, each containing *l* nucleotides. A line of best fit is calculated by linear regression and labelled the "local trend" within the segment. The detrended walk $y_l(i)$ is the difference between the original walk and the local trend. The squared detrended fluctuation is then defined as sum of squares of the detrended walks divided by the sequence length.

$$F_d^2(l) = \frac{1}{N} \sum_{i=1}^{N} y_l(i)^2. \tag{3}$$

To apply the method, one then assumes the sizes of the squared fluctuations follow a power law

$$F_d^2(l) \propto l^{\alpha} \tag{4}$$

where the exponent α can be found from the slope of a double-log plot. For random sequences, $\alpha = \frac{1}{2}$, but genomic sequences have larger values for α (Peng *et al*, 1994).

autocorrelation functions The auto-correlation function $C(r)$ was calculated following Messer *et al* (2005a,b). For a genomic sequence of length *L* and a sequence distance *r*

$$C(r) = \sum_{i}^{L} \sum_{n \in \{A,C,G,T\}} [P(a_i = a_{i+r} | a_i = n) - P(a_i = n)^2] \tag{5}$$

where $P(a_i = n)$ is the probability to encounter base n at position i and $P(a_i = a_{i+r})$ is the probability to find the same base in distance r from i.

Genomic data Bacterial genomes were taken from http://www.ncbi.nlm.nih.gov/genomes/static/eub.html

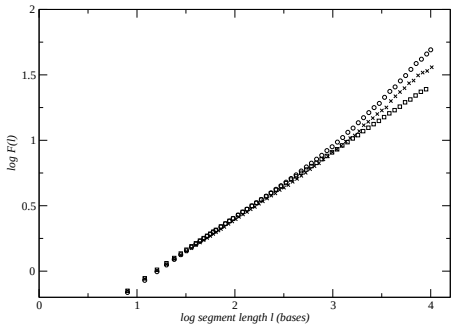

Fig. 1. Detrended fluctuation analysis. Bottom line (squares) from a random sequence, the middle line (crosses) from a simulated sequence as described in text and the top line (circles) from *E. coli*.

3 RESULTS

Initial Simulations

As listed under Methods, the code has seven adjustable probabilities, as well as initial conditions, the number of steps and the possibility to control the frequencies with which specific codons are introduced. A series of initial simulations were run to establish ranges for parameters, whether parameters could be combined and also to test the literature measures of genome properties on both simulated and real genomes.

Measures of Genome Properties

Root sequence effective length, L_r L_r depends on the word size k and should more correctly be written as $L_r(k)$. However, as has been reported, there is no additional information from different k values since one can fit data to a form $log_{10}(L_r(k)) = ak + b$ for some regression constants a and b (Chen *et al*, 2005b). A single value $k = 4$ was used for all calculations below. This was chosen because it is removes any susceptibility to patterns of length 3 (codon size) and avoids the problem that when the word size is too large, the statistics suffer due to noise (too few observations).

Detrended fluctuation analysis As originally described, this measure will reflect long-range order in a sequence (Peng *et al*, 1994). Fluctuations are measured, corrected for a baseline drift and the exponent α in equation 4 calculated from regression in a double-log plot. For a random sequence, $\alpha = 0.5$. This was easily confirmed as shown in Figure 1 (open squares). When long range order is present, one expects $\alpha > 0.5$. Unfortunately, there is no clear single slope as shown by the data from the genome of *E. coli* in Figure 1 (open circles). Instead, there is a change of slope around segment size of 10^2 to 10^3 as has been previously reported (Buldyrev *et al*, 1995). This is also the case with simulated sequences. The crosses in Figure 1 show α calculated from a run of 0.5×10^6 steps, segment length: 10 000, $\sigma_{segmentlength}$: 1 deletion 3: 0.199, mutation: 0.8, no triplet insertion and the only growth via segmental duplication.

Both real and simulated genomes always show the behaviour that there is an initial slope of $\alpha = 0.5$ followed by some larger value usually near $\alpha \simeq 0.7$. Rather than split the curves by hand, the second slope was obtained automatically and used in the comparisons below. First derivative of a curve was calculated numerically. This derivative curve can then be split into two parts, a and b.

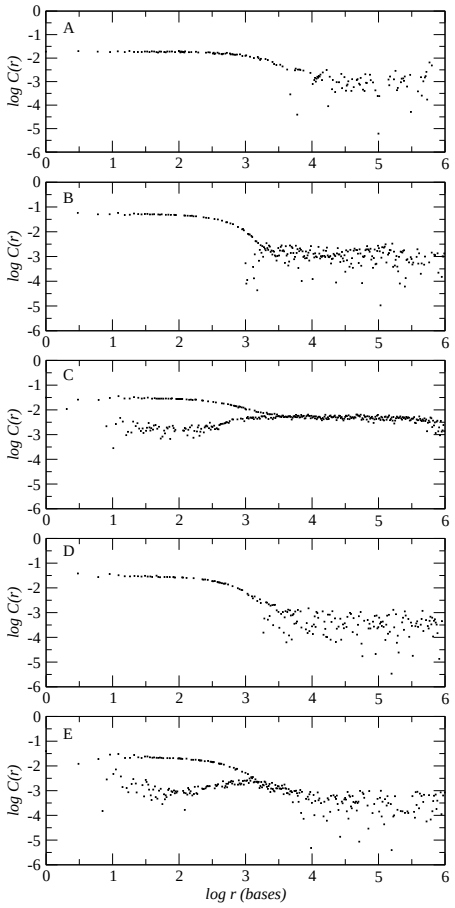

Fig. 2. Auto-correlation functions calculated from the (A) modelled sequence and genomic sequences (B) *Halobacterium sp*, (C) *Clostridium perfringens*, (D) *Mycobacterium tuberculosis H37Rv*, and (E) *Pyrococcus abyssi*

Within each part, the standard deviations, σ_a and σ_b were calculated. The splitting point was chosen so as to minimize the sum $\sigma_a + \sigma_b$ and the reported slope, α is from the steeper part of the curve at $log_{10} = 2.62$ or 417 bases.

Auto-correlation A simple autocorrelation function has been proposed as a measure of order in genomes and Figure 2 shows an analysis of an example simulation (parameters as described for Figure 1 and four genomes using the function as described by Messer *et al* (2005a,b). Unfortunately, not even a terminal optimist would attempt to fit a simple decay to these plots. This is not an implementation problem. One can see the same results from the authors' web server (Messer *et al*, 2006). With this result, we did not attempt any more quantitative analysis of autocorrelation functions.

Simulation Conditions

The first simulations served only to define the ranges of parameters which were possibly compatible with real genomes. Next, one could see which parameters had similar effects and could be ignored or combined. Finally, it was necessary to restrict the type of simulation conditions.

Table 1. Simulation parameter ranges

parameter	values	N_{values}
growth target (bases $\times 10^6$)	0.2, 0.4, 0.6, 0.8, 1.0	5
p(mutation)	0, 0.2, 0.4, 0.6, 0.8	5
χ_{sd}	0, 0.2, 0.4, 0.6, 0.8, 1.0	6
segment length (bases $\times 10^5$)	1, 2, 3, 4, 5, 6, 7, 8	8
number steps ($\times 10^6$)	0.5, 1.0	2
codon distribution	equi-dist, *E. coli*, emph *E. coli*	3
Initial sequence length (bases)	320 000	
Initial number of coding regions	300	
Initial length of coding regions (bases)	1000	
Initial Number of non-coding regions	10	
Initial length of coding regions (bases)	2000	

p(mutation) refers to the probability of a mutation event in a single step. equi-dist refers to a flat distribution of codons and *E. coli* to the emphasised bias due to *E.coli*.

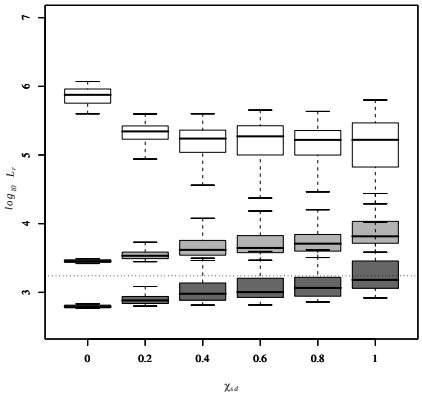

Fig. 3. Calculated effective root sequence length, L_r as a function of χ_{sd}. Top set of points (open boxes): no bias in codon selection. Middle (light grey) boxed points: *E. coli* codon bias. Bottom (dark grey) boxed points: emphasised codon bias. Each box shows 50% of the results range. Whiskers show the full range of values across all simulations. The dotted line marks L_r calculated from the bacterial genome set.

The first restriction was built into the model as described in the introduction. Single base pair operations were only allowed outside of coding regions. Only triplet operations were allowed within coding regions. After duplication, the duplicated segment could only be inserted into a non-coding region. After the initial simulations, more restrictions were added. The simulations were only to be of genome growth rather than steady state. This had an important consequence. Within the parameter ranges tested, single site operations had only a small effect. We then settled on testing a mixture of codon operations and duplications. Although the size of duplicated segments spanned a range from 1 to 8×10^5, the standard deviation was set to 1 base pair. The small variation served only to avoid any patterns due to a fixed segment size. All calculations began with the same sequence of 320 000 bases which was generated from 300 coding regions each of 1 000 bases and 10 non-coding regions of 2 000 bases length.

The set of parameters was reduced by taking advantage of those which were important, but correlated. Immediately, it was clear that if one is concentrating on genome growth, the dominant processes are triplet insertion and segmental duplication. A strategy was adopted to make the number of simulations tractable while mapping out the most important effects of the parameters. Each simulation had a target amount of growth ranging from 0.2 to 1×10^6 extra bases. Each sequence then was given a number of steps (0.5 or 1×10^6) to reach this target and a ratio of growth to deletion move probabilities calculated. The contribution of segmental duplication was labelled χ_{sd}, so if $\chi_{sd} = 0$, then there would be no segmental duplication. If $\chi_{sd} = 1$, there would be no insertion of bases. The probabilities of insertion and segmental duplication were scaled to account for their different sizes. For example, if $\chi_{sd} = 0.5$ the contribution to growth should be equal from the different mechanisms. If the segment size was 30 000, the probability of segmental duplication events would be 10^4 times less likely than codon insertion.

Finally, all simulations were run with either an even codon probability, probabilities taken from *E. coli* statistics (Nakamura *et al*, 1998) or the emphasised probabilities as described above. This led to a total of 7 200 calculations as listed in Table 1.

Simulation results

Some amount of information will be lost when summarizing >7 000 of simulations, but it is not difficult to distill the most important trends. Figure 3 shows the effective root sequence length, L_r as a function of χ_{sd}. The dotted line shows the single value calculated from the set of bacterial genomes ($log_{10}(L_r) = 3.2 \pm 0.2$ base pairs) from equation 1 and quoted $\pm$ one standard deviation. L_r for a random sequence is not shown as it is a quantity which simply scales with sequence length. The independent variable in the figure is χ_{sd}, but the box and whisker representation shows the range due to all the parameters. The median is marked at the centre of each data set. The box spans 50% of the results and the whiskers show the full range of values achieved by the parameters ranges as shown in Table 1.

The information can be summarized. The top series of points (all codons equally likely) do not come near the biological value, but as the rate of segmental duplication increases (right of figure), L_r moves in the direction of the biological value (down). The effect of a small (*E. Coli*) bias in codon selection is shown by the grey boxes. Even with no segmental duplication (left), L_r is not far from the biological value. Adding segmental duplication increases the effective root sequence length. Finally, using the emphasised codon bias (bottom points in dark grey) provides too small a value for L_r, but this reaches the biological value when segmental duplication increases. The plots should be interpreted with some care as they are really cross sections through the possible results. The whiskers show the full range of results seen by varying all parameters. For example, it is tempting to say that with an emphasised codon bias, the biological values are reproduced by very frequent segmental duplications ($\chi_{sd} = 1$). Looking at the other points shows that it is possible to reach the observed L_r in the range $0.4 \leq \chi_{sd} \leq 1$.

Figure 4 shows how the balance of the two growth mechanisms affects α, the exponent in the power law. The lower dashed line shows $\alpha = 0.5$, the value seen in a random sequence whereas the

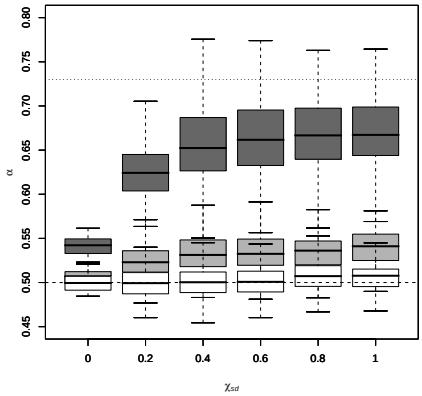

Fig. 4. Calculated α as a function of χ_{sd}. Bottom set of points (open boxes): no bias in codon selection. Middle (light grey) boxed points: *E. coli* codon bias. Top (dark grey) boxed points: emphasised codon bias. Each box shows 50% of the results range. Whiskers show the full range of values across all simulations. Random sequences (dashed line) show a power law coefficient of $\alpha = 0.5$. The dotted line marks α calculated from the bacterial genome set.

upper line near $\alpha = 0.73$ shows the value calculated from bacterial genomes and similar to the literature value (Peng *et al*, 1994). The only calculations to come near the experimental value are from simulations with an emphasised codon bias (dark grey boxes).

The effect of the other parameters was crudely checked using Kendall's rank correlation coefficient, but none are as obviously important as the codon bias or mechanism of genome growth. What constitutes sensible parameters is also a rather arbitrary decision. It is also clear that it is possible to generate almost any amount of order or disorder by manipulating the parameters. For example, one could simply use single base pair insertion and effectively generate random sequences. Hopefully the parameter ranges in Table 1 span more than a reasonable range.

4 DISCUSSION

A simple model cannot faithfully reproduce the properties of real sequences, but it should reproduce the trends associated with each evolutionary mechanism. We concentrate on the segmental duplication and codon bias, since these affect the results in a way that lends itself to comparison with other groups' simulations. Figure 3 shows the clear effect segmental duplication has on the effective root sequence length (L_r). Because this measure is apparently almost universal to genomes, this is one quantity that a simulation should be able to reproduce. Using an emphasised *E. coli* codon bias, one needs a substantial amount of segmental duplications to approach the experimental value. The trends are in agreement with previous work, where an even simpler model was used to suggest that segmental duplication alone could be responsible for this property in genomes (Hsieh *et al*, 2003; Chen *et al*, 2005b)

Varying the rate of segmental duplication also changes the power law exponent, α, as shown in Figure 4. The plot can be interpreted following Peng *et al* (1992; 1994). In a random sequence, one would would see $\alpha \approx 1/2$. Larger values for α are characteristic of longer range correlations. In the Figure 4, the exponent increases as the relative rate of segmental duplication increases (right hand side of plot). This may not be surprising, since copying pieces of a sequence into a different location copies all the properties of the source string and will look like long range order.

The effect of codon bias is perhaps more interesting because it is dramatic, but less well studied. Changing from unbiased codons to a distribution from *E. coli* and then to an emphasised distribution in Figure 3 moves the bars up in a similar manner to changing the relative rate of duplications. The effect on the power law exponent α (equation 4) in Figure 4 is equally dramatic. Codon bias, at least as implemented here, seems to be as important as segmental duplication.

The raises the question as to whether the bias in codon selection leads to some bias in the genome. The answer is that the genome composition will always be less biased than the selection of codons. One inserts codons with a certain frequency, but the single base mutations dilute this. This can be confirmed by sampling the final strings. With no bias, the composition {[A], [C], [G], [T]} is {0.25, 0.25, 0.25, 0.25}. With *E. coli* bias, one sees {0.26, 0.24, 0.26, 0.25} and with emphasised bias, one typically sees {0.27, 0.22, 0.27, 0.24}. The exact values vary from run to run, but the final base composition did not vary by more than $\pm\, 0.01$.

The results can be compared with other simulations that were intended to model biological mechanisms. Local duplication was one of the core elements of the model of Messer *et al* (2005a; 2005b). There, copied bases were inserted next to their source. This had the advantage that the model lent itself to an elegant analytical analysis as well as numerical simulation. It also fit well to the measure of order used by the authors. Adjacent regions of sequence will be highly correlated and this correlation will decay due to the other processes acting on the sequence. Since their model had no specific triplet (codon) operations it yielded a smooth auto-correlation decay function. They also noted that qualitatively similar behaviour would be seen if whole segments were duplicated (Messer *et al*, 2005a,b). In this work, duplicated segments could be placed anywhere and changes within coding regions were dominated by triplet mechanisms. This means that one will not see a simple decay and the auto-correlation function is not a good measure. This is much of the reason for the tremendous noise seen in Figure 2. One would certainly not want to change this part of the model as it is a property seen in real bacterial genomes (Figure 2 B).

The expansion-modification system is very similar in terms of the proposed mechanisms (Li, W. , 1991). The may be a bit unexpected since unlike Messer *et al* (2005a; 2005b), the results have been interpreted in terms of $1/f$ noise which one would see as a longer range property (Li *et al*, 1994, 1995).

The simulations of Hsieh *et al* 2003 concentrated on segmental duplication and found it essential for reproducing a reasonable effective root sequence length L_r. Obviously, our results do not contradict this. We would however claim that if one is going to interpret patterns in genomic sequences in terms of physical mechanisms, one cannot ignore a bias in the way codons are introduced. Because one is changing three bases at a time, this mechanism will not suit conventional auto-correlation analysis as used by others (Messer *et al*, 2005a,b), but nor do real genomes (Figure 2).

One should look at the results and possible parameters more generally. Figures 3 and 4 show a range of results in terms of quartiles,

but the spread of results does not come from statistical error. This is too small to show on these plots. The range represents the values seen by varying the other parameters as shown in Table 1. This means that the parameters such as mean segment length also affect the results. Varying even more parameters such as the rates of single site mutations, deletions or insertions had similar effects (results not shown). This means that one is dealing with a large set of distinctly non-orthogonal variables. A change in one may be compensated for a change in another.

The best way to reconcile the mechanisms is to begin with the simpler models such as expansion-randomization (Messer *et al*, 2005a,b) or expansion-modification (Li, W. , 1991). The expansions (through duplication) increase local correlation while the randomizing mechanisms dilute this and lead to a smooth decay. More generally, one could talk in terms of a source of order and sinks which remove it. Depending on one's model, the source may be segmental duplication (Hsieh *et al*, 2003), local duplication (Li, W. , 1991; Messer *et al*, 2005a,b) or a bias in codon selection. This order may be transported by the local expansion or transported over longer distances by a copying mechanism. Any model with these properties can probably reproduce many properties of genomes whether they be labelled order, correlation or information. In terms of biological mechanisms, the more imaginative could even invoke chemical justifications such as stabilization due to base stacking interactions de Souza and Anteneodo (1995).

So far, the model from this work has been compared to some simulations which were constructed to reproduce some specific properties. The model, however, fits into a spectrum ranging from coarse-grain to detailed approaches. At the simpler end, one essentially has a hidden Markov model which will reproduce built-in correlations (de Souza and Anteneodo, 1995). At the most realistic end of the spectrum, there is a system which should be detailed enough to regenerate specific phylogenies (Beiko and Charlebois, 2007). The model here is was built to be sufficient to capture some of the mechanisms which may lead to properties labelled as order, structure or information. It does deal with four base types, codons as a unit and has a minimal version of genes and non-coding regions. There are however biological processes which might be relevant, but were not included. There is no mechanism for deleting large pieces of sequence, although this would serve a source of order if applied to less ordered regions. There is no mechanism for speeding mutation after a gene is copied and may be redundant. Segmental duplications were a major factor, but there were no sources of exogenous sequence. The model has no site or region specific rates of change, although this is routinely discussed in phylogenetic studies, because of the effect on molecular clocks Brocchieri (2001).

Continuing in this speculative vein, there is an element of time which is completely ignored. The model begins with a small genome which grows by a series of random events, all with fixed probabilities. It is likely that in real evolution, relative rates have not been fixed. Perhaps duplications of segments were rather common in early small genomes, whereas smaller changes became more important later. Perhaps one wants a model closer to punctuated evolution. The obvious problem with this speculation is that each option adds literally another dimension of adjustable parameters.

A more fruitful area of progress may not be in the model for genome simulation, but in the properties measured. One could summaries the analysis of genomes and their statistics by saying that most workers are looking at some property related to order, whether it be fluctuations, long or short range correlations or a concept from information theory. From the results here, naïvely applying a physics-based measure does not always give meaningful results. An auto-correlation function applied to a real genome is disastrous (Figure 2) and the results of detrended fluctuation analysis on biological data do not yield a simple single slope (Buldyrev *et al*, 1995).

It may be that the most interesting result here is to numerically show how different processes lead to similar results and to suggest that many literature models are essentially complementary. The implication is that if one wants to speculate further about the exact mechanisms which led to more modern genomes, one may have to find more discriminative statistical properties.

ACKNOWLEDGEMENTS

We thank the German Academic Exchange Service (DAAD) and the National Science Council (ROC) of Taiwan for financial support.

REFERENCES

Beiko, R.G. and Charlebois, R.L. (2007) A simulation test bed for hypotheses of genome evolution, *Bioinformatics*, **23**, 825-831.

Brocchieri, L. (2001) Phylogenetic inferences from molecular sequences: review and critique, *Theor. Popul. Biol.*, **59**, 27-40.

Bromham, L. and Penny, D (2004) The Modern Molecular Clock, *Nat. Rev. Genet.*, **4**, 216-224.

Buldyrev, S.V., Goldberger, A.L., Havlin, S., Mantegna, R.N., Matsa, M.E., Penk, C.-K. and Stanley, H.E. (1995) Long-range correlation properties of coding and non-coding DNA sequences: GenBank analysis, *Phys. Rev. E*, **51**, 5084-5091.

Chen, T.Y., Hsieh, L.C., Chang, C.H. Luo, L.F. and Lee, H.C. (2004) *Int. J. Mod. Phys. B*, **18**, 2448-2454.

Chen, H.D., Chang, C.H., Hsieh, L.C. and Lee, H.C. (2005a) Divergence and Shannon Information in Genomes, *Phys. Rev. Lett.*, **94**, 178103.

Chen, T.Y., Hsieh, L.C., Lee, H.C. (2005b) Shannon information and self-similarity in whole genomes. *Comput. Phys. Commun*, **169**, 218-221.

Fadiel, A., Lithwick, S. and Naftolin, F. (1995) The influence of environmental adaptation on bacterial genome structure, *Lett. Appl. Microbiol.* **40**, 12-18.

Foerstner, K.U. ,von Mering,C., Hooper, S.C. and Bork, P. (2005) Environments shape the nucleotide composition of genomes, *EMBO Rep.*, **6**, 1208-1213.

Gillespie, D.T. (1977) Exact Stochastic Simulation of Coupled Reactions, *J. Phys. Chem.*, **81**, 2340-2361.

Havlin, S., Blumberg-Selinger, R., Schwartz, M., Stanley, H.E. and Bunde, A. (1988) Random multiplicative processes and transport in structures with correlated spatial disorder, *Phys. Rev. Lett.*, **61**, 1438-1441.

Hsieh, L-C., Luo, L., Ji, F. and Lee, H.C. (2003) *Phys. Rev. Lett.*, **90**, 018101.

Li,W., Marr, T.G. and Kaneko, K. (1994) Understanding long-range correlations in DNA-sequences (1994) *Physica D* **75**, 392-416.

Li,W., Marr, T.G. and Kaneko, K. (1995) Erratum to Understanding long-range correlations in DNA-sequences (1995) *Physica D* **82**, 217.

Li, W. (1991) Expansion-modification systems: A model for spatial $1/f$ spectra (1991) *Phys. Rev. A*, **43**, 5240-5260.

Messer, P.W., Lässig, M. and Arndt, P.F. (2005a) Universality of long-range correlations in expansion randomization systems, *J. Stat. Mech.*, P10004.

Messer, P.W., Arndt, P.F. and Lässig, M. (2005b) Solvable sequence evolution models and genomic correlations *Phys. Rev. Lett.* **94** 138103.

Messer, P.W. and Arndt, P.F. (2006) CorGen-measuring and generating long-range correlations for DNA sequence analysis, *Nucleic Acids Res.*, **34**, W692-W695.

Peng, C.K. Buldyrev, S.V., Goldberger, A.L., Havlin, S., Sciortino, F., Simons, M., Stanley, H.E. (1992) Long-range correlations in nucleotide sequences, *Nature* , **356**, 168-70.

Peng, C.K., Buldyrev, S.V., Havlin, S., Simons, M., Stanley, H.E., Goldberger, A.L. (1994) Mosaic organization of DNA nucleotides, *Phys Rev E*, **49**, 1685-1689.

Nakamura, Y., Gojobori, T. and Ikemura, T (1998) Codon usage tabulated from the international DNA sequence databases, *Nucl. Acids Res.* **26** 334.

Rosner, B. (2000) *Fundamentals of biostatistics, 5th ed*, Duxbury.

de Souza, A.M.C. and Anteneodo, C. (1995) A Model for Nucleotide Sequences, *Biophys. J.*, **69**, 1708 - 1711.

Tai, Y.Y., Li, P.C., Tseng, H.C. (2006) A two-dimensional modified Levy-walk model for the DNA sequences, *Physica A*, **369**, 688-698.

Voss, R. (1992) Evolution of long-range fractal correlations and 1/f noise in DNA-base sequences, *Phys. Rev. Lett.* **68**, 3805-3808.