Models of Genome Evolution

Yi Zhou[1] and Bud Mishra[2,3]

[1] Biology Department, New York University, 100 Washington Square East, New York, NY 10003, USA, joey@cs.nyu.edu
[2] Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10012, USA, mishra@cs.nyu.edu
[3] Watson School of Biological Sciences, Cold Spring Harbor Laboratory, 1 Bungtown Rd., Cold Spring Harbor, NY 11724, USA

Summary. The evolutionary theory, "evolution by duplication", originally proposed by Susumu Ohno in 1970, can now be verified with the available genome sequences.

Recently, several mathematical models have been proposed to explain the topology of protein interaction networks that have also implemented the idea of "evolution by duplication". The power law distribution with its "hubby" topology (e.g., P53 was shown to interact with an unusually large number of other proteins) can be explained if one makes the following assumption: new proteins, which are duplicates of older proteins, have a propensity to interact only with the same proteins as their evolutionary predecessors. Since protein interaction networks, as well as other higher-level cellular processes, are encoded in genomic sequences, the evolutionary structure, topology and statistics of many biological objects (pathways, phylogeny, symbiotic relations, etc.) are rooted in the evolution dynamics of the genome sequences.

Susumu Ohno's hypothesis can be tested '*in silico*' using Polya's Urn model. In our model, each basic DNA sequence change is modeled using several probability distribution functions. The functions can decide the insertion/deletion positions of the DNA fragments, the copy numbers of the inserted fragments, and the sequences of the inserted/deleted pieces. Moreover, those functions can be interdependent. A mathematically tractable model can be created with a directed graph representation. Such graphs are Eulerian and each possible Eulerian path encodes a genome. Every "genome duplication" event evolves these Eulerian graphs, and the probability distributions and their dynamics themselves give rise to many intriguing and elegant mathematical problems.

In this paper, we explore and survey these connections between biology, mathematics and computer science in order to reveal simple, and yet deep models of life itself.

1 Introduction

The genome of an organism is a collection of its genes, encoded by four chemical *bases* in its DNA (DeoxyriboNucleic Acid), and forms the genetic core of a cell. The genes ultimately encode for the proteins (chains of amino acids) and in turn, the genes themselves are regulated by transcription factors and other operons, many of which are proteins. The sequences of amino acids, specified by the DNA through transcription and translation processes, determine the three-dimensional structure and biochemical properties of the proteins as well as the nature of their interactions. Furthermore, mRNA stability, protein degradation, post-translational modifications, and many other biochemical processes tightly regulate the time-constants involved in the resulting biochemical machinery. Proteins also associate in complexes to form *dimers* (pairs of proteins), *trimers* (triplets), and *multimers*. An *isoform* of a protein is a slightly different protein with a closely related sequence, and often shares similar functional properties, e.g., enzymatic reactions, but is regulated differently. However, we know that this complex machinery of life has evolved over several billion years through random mutations and re-arrangements of the underlying genomes while being shaped by the selection processes. We ask the following questions: Are the processes altering, experimenting and correcting the genomes completely untamed and haphazard? If not, what signatures have they left on the genomes, proteomes, pathways, organs and organisms? What structures have they imposed on these biological elements and their interactions? We posit that a better understanding of biology is hinged on a deep information-theoretic study of evolving genomes and their roles in governing metabolic and regulatory pathways.

We begin with the following account: Genomes are not static collections of DNA materials. Various biochemical and cellular processes—including point mutation, recombination, gene conversion, replication slippage, DNA repair, translocation, imprinting, and horizontal transfer—constantly act on genomes and drive the genomes to evolve dynamically. These alterations in the genomic sequences can further lead to the corresponding changes in the higher-level cellular information (transcriptomes, proteomes and interactomes), and are crucial in explaining the myriad of biological phenomena in the higher-level cellular processes. However, until recently, the lack of sufficient historical data and the complexity of biological processes involved have hampered the development of a rigorous, faithful, and yet simple abstract model for genome evolution.

Present genomes can be viewed as a snapshot of an ongoing genome evolution process. Although it would be ideal, it is usually impossible to base genome evolution studies on ancient genome samples. Fortunately, various historical evolutionary events leave their "signatures" in the present sequences, which can be deciphered by statistical analyses on a family of genomes that are currently available. With the development of high throughput experimental technology, the flow of information at different levels of biology (genome,

proteome, transcriptome, and interactome) is increasing dramatically. By analyzing and comparing this data, we are now able to look for the structure of cellular processes and the dynamics of the evolution process driving it.

A survey of the literature reveals many interesting statistical analyses of various kinds on genomic and proteomic data. Among the large collection of results, it is worthwhile to note that many interesting statistical characteristics are shared by data from all organisms, from different cellular processes, as well as at various scales. For example, research during the last decade reveals long range correlation (LRC) between single nucleotides in the genomic sequences of various species from different kingdoms [1][20], and in different regions of the genomic sequences. Furthermore, the LRC is persistent in all the genomic sequences examined. This indicates that on single-nucleotide level, genomes have evolved independently to share a common scale-free global structure. On a slightly larger scale than single nucleotides, the short words in DNA sequences (mers, oligonucleotides) and protein sequences (short peptides) also seem to display similar generic statistical properties. The frequency distributions of the short words in the genomic and proteomic sequences from various organisms are found to follow a power-law [18][14], a feature often found in linguistic studies. Those general properties are further reflected in the higher-level cellular processes. As the large-scale metabolism networks and protein interaction networks in some model organisms become available, e.g. metabolic networks in *E. coli* [8] or protein interaction networks in *S. cerevisiae* [3] and *H. pylori* [11], the topology of those networks is analyzed [9][19] and is found to be characteristic of a group of graphs known as scale-free networks [4]. Scale-free networks are characterized by their "hubby" structures associated with a power-law distribution of their connectivities, and can be created by an evolution process following a"rich gets richer" rule. All those statistical features—the positive correlation between single nucleotides, the over-representation of high-frequency words in genomes and proteomes, the "hubbiness" due to highly-connected nodes in the protein and genetic networks—can be viewed as the different aspects of an underlying generic structure. (See Genomic data Analysis section for more examples.) Different organisms preserve a common structure in their cellular processes despite drastically different evolutionary environments. This common structure may reflect the most fundamental processes in biology.

The positive feedback mechanism suggested by the highly-correlated structures found in various data is reminiscent of the "evolution by duplication" theory originally proposed by S. Ohno in 1970's [15]. Based on this theory, we develop a mathematical model to explain the observations in our mer-frequency distribution analysis. The model is an extension of Polya's urn model [13], and considers genome evolution as a stochastic process with three main events: *substitution*, *deletion*, and *duplication*. We study a simpler version of the model in numerical simulation as well as a more realistic, thus more complicated, version of the model in a large-scale *in silico* evolution simulation. The simple model fits the real-world data for mer-frequency dis-

tributions. These results suggest that despite the highly diversified evolutionary environment for different organisms, the essential composition of the evolutionary dynamics is commonly shared. A simple stochastic process (*substitution*, *deletion*, and *duplication*) can describe the recurrent pattern in the statistical signatures of different organisms. The model is extremely intriguing as it suggests that all the complexities found in life can be the result of a simple stochastic evolutionary process.

2 Evolution by duplication

Susumu Ohno proposed an evolutionary theory "evolution by duplication" [15]. Although not explicitly stated, his theory suggested a "rich gets richer" rule in genome evolution. The theory argues that the evolutionary advantage of evolution by duplication lies in the promise that with an extra copy, the selection pressure on the gene is somewhat relaxed. Since the original function can be maintained efficiently by either copy of the duplicated gene, the other copy can undergo various modifications, increasing the chance of the organism obtaining a new advantageous gene. Gene duplication can speed up the search for higher fitness in various ways [23]: it can adjust gene dosage; attain a permanent heterozygous advantage by incorporating two former alleles into the genome, allow more specialized functions by differential regulation of the duplicated genes, or create a new gene with a diverged function.

The duplication process mentioned in the theory can be well explained by molecular biology. There are various molecular mechanisms that can cause DNA duplication of different size ranges. For example, during DNA replication, replication slippage [6] can introduce small insertions or deletions locally, when the newly replicated DNA fragments misalign to the template. The misalignment, usually triggered by tandem repeats or secondary structure in the template DNA strand, causes the DNA polymerase to pause, dissociate, and continue an erroneous strand extension after re-association. During meiotic cell division, recombinations between two DNA molecules occur through cross-overs between corresponding homologous regions. Unequal cross-overs between two DNA molecules bearing successive repeated fragments will result in the duplication or deletion of the repeated units in the daughter cells [23]. Another process that can introduce duplications or deletions of relatively large sizes and globally in the genome involves mobile DNA elements (insertion elements, transposons, and retrotransposons) [23]. The mobile elements can be either excised or copied from their original positions, and subsequently inserted elsewhere in the genome where target sequences can be found. The frequencies and sizes of the deletions or insertions vary with specific elements. Since the target sequences are widely distributed in the whole genome, the mobile elements can essentially affect sequence changes on the whole-genome range. In summary, duplications and deletions in genomes can be fully justified by the well-known molecular mechanisms.

However, the duplication dynamics do not proceed completely unopposed—the cell also possesses DNA repair machineries to counteract the changes made in the sequences and prevent genomes from changing too rapidly [7]. For example, the mismatch repair (MMR) mechanism is mainly responsible for correcting most of the deletions and insertions of various sizes. Therefore, duplications, deletions, and other changes in genomes are the results of the interactions between the molecular mechanisms leading to genomic sequence changes and the surveillance system of the cell.

If we assume that the target gene of every duplication is randomly chosen from the genes that are already in the genome, then we have a realization of Polya's Urn model [13]. Therefore, under the "evolution by duplication" theory, genome evolution can be viewed as a stochastic duplication process that can lead to a highly correlated structure with over-abundance of some elements.

Several mathematical models recently proposed to explain the topology of protein interaction networks have also implemented the idea of "evolution by duplication" [2][17][10]. The power-law distribution with its "hubby" topology (e.g., P53 was shown to interact with an unusually large number of other proteins [12]) can be explained if one makes the assumption that new proteins which are duplicates of older proteins have a propensity to interact only with the same proteins as their evolutionary predecessors. Since the protein interaction networks, as well as other higher-level cellular processes, are encoded in genomic sequences, the evolution of their topology is rooted in the genomic sequence changes. Therefore, we believe that a more general model of "evolution by duplication" at genomic level should explain the common pattern observed at various scales and different cellular information levels, and may be exploited prudently in the design of better bioinformatics algorithms.

3 Genomic data analysis

Large-scale genomic data analysis is the essential starting point in the search for the main players during genome evolution. Previous researches have provided statistical evidence favoring a general genome evolution dynamic, "evolution by duplication". Here, we report further discoveries that support such a hypothesis. We have examined the statistical properties of the distribution of short words in various whole genomes and proteomes. Our results confirm and extend the previous conclusion that there is an over-representation of high-frequency words in all the sequences studied. Furthermore, our analysis of the distribution of the end-points of putative large segmental duplications in human genome provides convincing evidence that duplications tend to occur more often around the prior duplication sites. In another words, the already duplicated segments are more likely to be duplicated again, thus the already over-represented segments tend to be more over-represented, while the other segments are more likely to be suppressed. The duplication dynamics im-

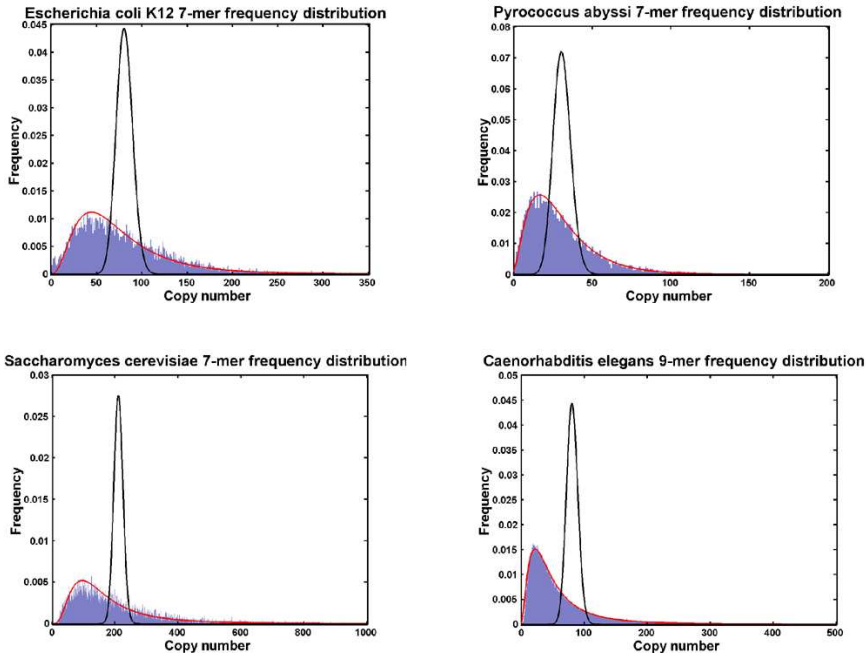

Fig. 1. Mer-frequency distribution in some genomes examined. The plots shows the non-overlapping mer frequency distribution (blue bars) in the genomes of a eubacteria (*E. coli* K12), an archea (*P. abyssi*) and two eukaryote (*S. cerevisiae* and *C. elegans*). When compared with the expected distribution from a random sequence of the same length (black line), the distributions from real sequences consistently show an over-representation of high-copy mers. Our simulation results (red line) from the simple graph model closely fit the "real" mer frequency distribution. Given that we only have two free parameters (q = single point mutation probability and p_1/p_0 = ratio of probabilities of duplication over deletion, see below) in the model, the data-fitting is extremely convincing.

plied by the end-point distribution analysis may explain our observations on protein domain family sizes, which follow a power-law distribution and are characterized by an over-representation of larger families.

3.1 Mer-frequency distribution analysis

To study the statistics of short words in different whole genomic sequences, we performed a large-scale non-overlapping mer-frequency distribution analysis. The experiment was conducted on all the reasonable mer-sizes, covering almost all the presently available whole genomic sequences and including various organisms from all the kingdoms. To avoid the complication of inversions, we treated two inversely complimentary mers as one species. (For example,

5'-ATCG-3' and 5'-CGAT-3' are counted as one mer species, i.e., their frequencies are combined.) Therefore, for mer size l, there are $\frac{4^l}{2}$ species of l-mers. From our results, it is clear that the mer-frequency distributions from all the genomic data examined deviate from the random distribution (see Figure 1 for some of the results). Furthermore, they are all characterized by the same type of deviation—over-representation of high frequency mers.

We have also looked at the mer-frequency distribution in just the coding sequences, and the distribution of amino-acid word-frequency in the corresponding proteome sequences. (For a length of n, there are 20^n species of different amino-acid words.) Both results share the same type of deviation from the random distribution that is observed in the whole genomic sequences (data not shown).

A simple simulation of "evolution by duplication" was also performed, where a short random sequence (1000bp) was allowed to evolve to a final length of 500Kb by duplicating fragments randomly chosen from itself. The deviation in the mer-frequency distribution of the final sequence from a random sequence closely resembles the pattern seen in real genomes (see Figure 2). Therefore, the particular statistics of mers in genomes and short amino-acid words in proteomes can be simply due to the duplication processes during genome evolution.

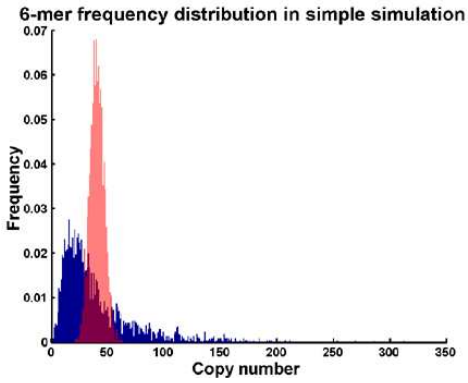

Fig. 2. The 6-mer frequency distribution of the resulting sequence of a simple "evolution by duplication" simulation. The initial condition is a random sequence of length 1000bp. The sequence is evolved through multiple iterations until it reaches a length of 500Kb. In each iteration, a fragment of length uniformly randomly distributed from 1 to 100bp is randomly chosen from the sequence, duplicated, and re-inserted randomly into the sequence. The blue bars in the plot show the 6-mer frequency distribution of the final sequence from the simulation. The red bars show the 6-mer frequency distribution of a random sequence of the same length.

3.2 Analysis on the end-points of potential large segmental duplication fragments in human genome

As mentioned above, the results of various statistical analyses on genomic data have suggested that there is a generic evolution dynamic dominated by duplication. To justify this hypothesis, it is important to study the dynamic of duplication processes. Although the exact molecular mechanisms that cause duplications are not fully understood, we can approach the problem indirectly by looking at the distribution of the most recent segmental duplications in genomes.

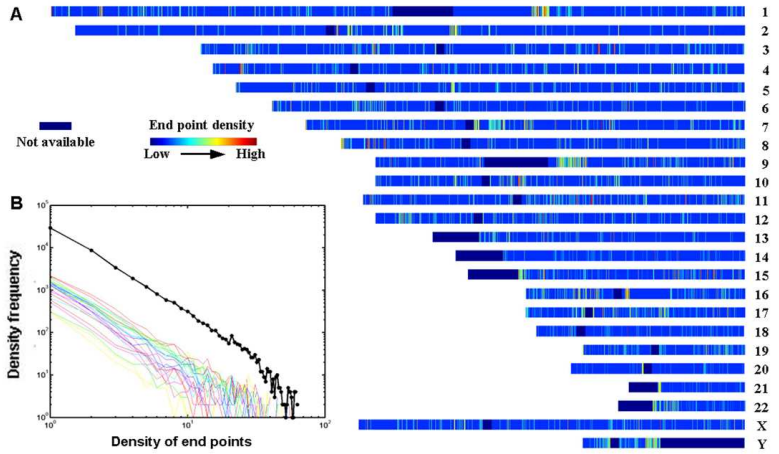

Fig. 3. The distribution of the potential duplication 'hot-spots' on the human genome. A. The distribution of the duplicated segment end-points on the chromosomes (over windows size of 1Kb). The 'hot-spot' density is color coded (see the color bar). The dark areas represent chromosomal regions where no reference sequences are available. There is a tendency for areas with high densities to cluster together on the chromosomes. B. The distribution of the end-point densities on a log-log scale. The X axis shows the number of end-points in non-overlapping windows of size 1Kb (density of end-points), starting from 1. The Y axis indicates the number of non-overlapping 1Kb windows containing a given number of end-points (density frequency). It is clear that the $Log_{10}(density of end-points)$ and $Log_{10}(density frequency)$ form a linear relationship, both in the whole-genome range (thick black points) and in individual chromosome range (multiple thin colored lines).

Recently, intensive large segmental duplications (both intra- and inter-chromosomal) have been reported in the assembled human genome [21], and the potential large duplicated regions (>500bp, >95% identity) have been mapped out in pairs under the standard sequence homology criteria [22].

Although the exact molecular mechanism is unclear, one could hypothesize that some of the processes involve single- or double-strand DNA breakage at the time of duplications, like homologous or heterologous recombination, and transposition. The end-points of the mapped-out duplicated segments are good candidate sites where such processes initiate or terminate at the time of duplications.

Since the end-points can be viewed as the signatures left by the duplication process over the genome evolutionary history, their distribution along the genome could reveal some dynamic features of the duplication process. To verify this hypothesis about the "density" distribution of the end-points, we first fragmented the genome into non-overlapping windows of a fixed size. The number of end-points covered by each window was treated as its local "density" over the corresponding genomic region. When we plotted the histogram of the end-point density over a chromosome or over the whole genome, we discovered another power-law distribution (Figure 3). This implies that duplications tend to happen more often at the previous duplication sites, which is driven by a positive feedback dynamic. To further verify this interpretation, we performed a correlation test (detrended fluctuation analysis) [5] on the series of densities along the genome in relation to the distances between them. The result of the test indicated a positive correlation between neighboring densities of end-points. Such positive correlation suggests that over the evolutionary history, consecutive segmental duplications occur favorably near or on some previously duplicated segments, and are absent elsewhere.

3.3 Protein domain family size distribution

In our mer-frequency distribution analysis, the chosen mer sizes ranges up to 12 nucleotides. In comparison to other functional elements in the genome, the mers are of the smallest scale. To check whether the generic structure observed on such a small scale also persists on a bigger scale, we studied the protein domain family size distribution.

The protein domain families in different organisms are extracted from the protein family database InterPro [16]. The sizes of the domains are mostly above 50 amino acids (150 nucleotides) — a bigger scale than that of the mer analysis. The analysis of domain families is based on sequence signature and homology. A family of protein domains found by this method can be viewed as a cluster of amino acid sequences from a proteome that share enough similarity with each other and have maintained their critical sequences. When the histograms of the sizes of those protein domain families from various organisms were plotted on a log-log scale, a linear relationship was observed in all cases, including in *E. coli* K12, *P. abyssi*, *S. cerevisiae*, *H. sapiens* (Figure 4). Therefore, the domain family size distribution, or more generally, the cluster size of homologous amino acid sequences, seems to follow a power-law distribution. Here, on this larger scale, we observed the same deviation pattern from the random distribution as in the small-scale mer distribution

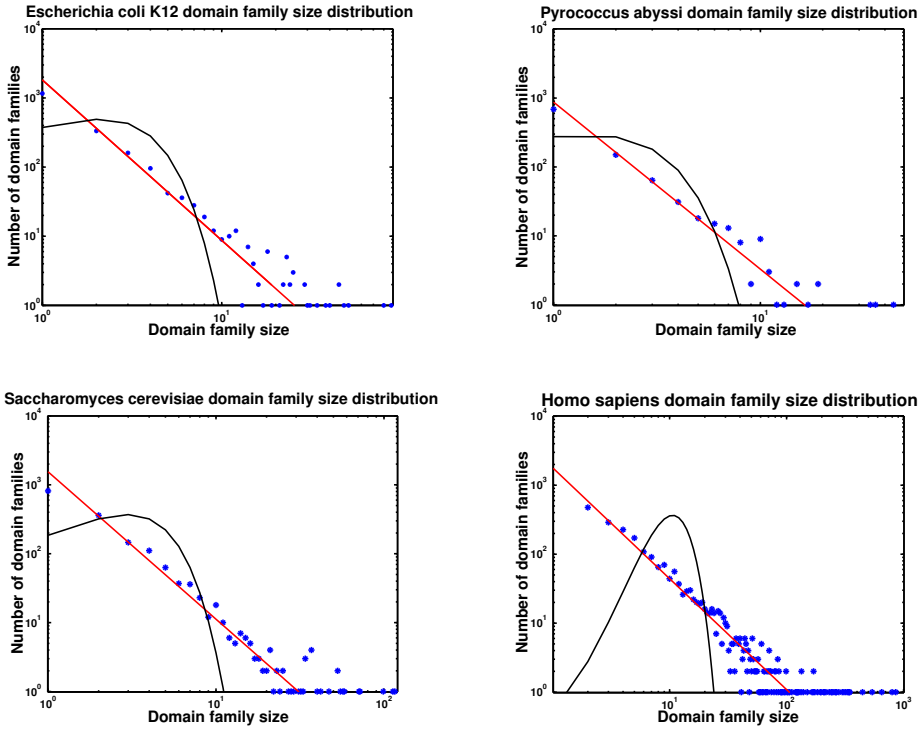

Fig. 4. Protein domain family size distribution in some genomes examined. The plots show the protein domain family size distribution in the corresponding proteomes (blue dots). The protein domain family data is extracted from InterPro database. The plots are on a log-log scale. The almost linear shape (red line) on such a plot indicates a power law relationship between a domain family size and the number of domain families of that size. Therefore, the protein domain family size distributions are also characterized by an over-representation of large size families when compared with uniformly random distributions (black curves).

analysis, which further confirms the effects of duplication processes during genome evolution.

4 Evolutionary models

The results from genomic data analysis at different scales, at different levels, and in different organisms repeatedly show the same pattern (over-representation of high-frequency elements). It consistently suggests a generic evolutionary dynamic involving positive feedback as postulated by S. Ohno's theory of "evolution by duplication." If we assume that the target fragment of every duplication is uniformly randomly chosen from the genome, then the

fragments that are already over-represented in the genome will have a higher probability of getting duplicated again. Similarly, the fragments that are initially under-represented in the genome will be further suppressed. To further verify the theory quantitatively, we develop a mathematical model based on the theory.

4.1 Graph model

We develop a Eulerian graph model to explain the frequency distribution of non-overlapping mers (all the different species of oligonucleotides of a particular size) in various genomic sequences. The model aims to capture the parsimonious processes needed to recover the dynamics involved in genome evolution. Yet, it preserves enough fidelity to validate biological reality. The processes included in the model are: *duplication*, *deletion*, and *substitution*. The parsimony of the model can be inferred from the fact that the omission of any of the three processes renders the model unsuitable to fit real genomic data. In the model, a genome is represented by a Eulerian graph. Each mer species of a particular length is represented by a node. Whenever two non-overlapping mers are immediately adjacent to each other in the genome, they are connected by an additional directed edge. Without loss of generality, the edges are always directed from the 5' end to the 3' end. Therefore, the number of directed edges from node i to node j ($k_{i,j}$) indicates how many times the i^{th} mer is immediately adjacent to the 5' end of the j^{th} mer. We use k_i to represent both the out-degree (${k_i}^{out}$) and the in-degree (${k_i}^{in}$) of the node i, since due to the Eulerian property of the graph, each node has identical in- and out-degrees, each being equal to the copy number of the corresponding mer in the genome. For mers of size l, and a genome of length L, the graph will have a total of $N = \frac{4^l}{2}$ nodes and $E = \frac{L}{l} = \sum_{i=1}^{N} k_i$ edges. Such graphs are Eulerian and each possible Eulerian path in the nontrivial (non-singleton) connected component encodes a genome. The genomes represented by the same graph share the same mer frequency distributions but not necessarily the same arrangement of mers.

The evolution of a genome is modeled as a stochastic evolution process of the graph that goes through multiple iterations. The model assumes that all the presently existing genomes originated from a proto-genome, which is very small and its has are randomly distributed mers. Thus, the initial graph is a random graph with a small average degree. In each iteration, one of the three possible processes occurs: *duplication* of a chosen mer (with probability p_1), *deletion* of a chosen mer (with probability p_0), or *substitution* of a chosen mer by another mer (with probability q) (Figure 5). Therefore,

$$p_1 + p_0 + q = 1 \tag{1}$$

To avoid extinction, we let $p_1 > p_0$. Biological processes that can cause duplications or deletions include homologous or heterologous recombination

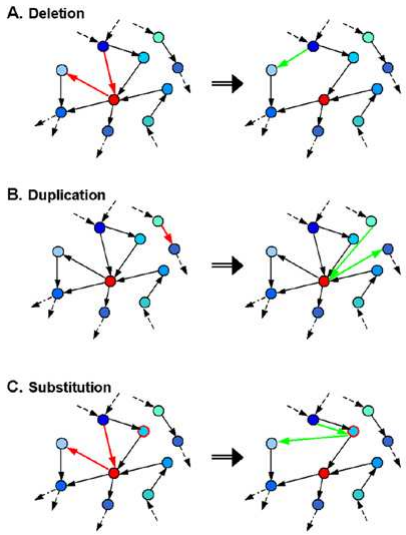

Fig. 5. The three processes during graph evolution: *deletion*, *duplication*, and *substitution*. In each process, the target node (indicated by red fill in the figure) is chosen with preference for nodes with larger degrees: If the i-th node has degree k_i, the probability of it being chosen is proportional to $\frac{k_i}{\sum_{i=1}^{N} k_i}$. For deletion process (**A**), one incoming edge and one outgoing edge of the target node are randomly chosen (red), and deleted from the graph. A new forward edge (green) is added from the root node of the deleted incoming edge to the head node of the deleted outgoing edge. For duplication (**B**), an edge is randomly selected from the graph (red), and deleted. Two forward edges (green) are added from the root node of the deleted edge to the target node, and from the target node to the head node of the deleted edge, respectively. For substitution (**C**), besides the target node as the substituted node, another node (indicated by red boundary) is randomly chosen from the graph uniformly as the substituting node. One incoming and one outgoing edge of the target node are randomly selected (red), and rewired to the selected substituting node (green). Note that all the processes during graph evolution keep the out-degree and in-degree of a node identical.

and DNA polymerase slippage. Substitutions can be caused by random point mutation. During graph evolution, let ${k_i}^t$ and E^t indicate the copy number of i^{th} mer and the total number of mers in the evolving genome at t^{th} iteration. If we assume that the target mers for any process is chosen uniformly randomly from the genome, then the probability of i^{th} mer species being chosen for a process in the next iteration is proportional to its frequency in the genome in the current iteration ($\propto \frac{{k_i}^t}{E^t}$). With these assumptions, we implemented the "rich gets richer" rule in dynamics. If, for simplicity, a mer chosen for substitution is assumed to change into any other mer with equal probability during

substitution, the conditional probabilities describing how the copy number of the i^{th} mer changes in each iteration can be expressed as follows[4]:

$$P({k_i}^t = n \mid {k_i}^{t-1} = n-1) = p_1 \frac{n-1}{E^{t-1}} + (1 - \frac{n-1}{E^{t-1}}) q \frac{1}{N-1}, \tag{2}$$

$$P({k_i}^t = n \mid {k_i}^{t-1} = n) = 1 - p_1 \frac{n}{E^{t-1}} - p_0 \frac{n}{E^{t-1}} - q \frac{n}{E^{t-1}} - (1 - \frac{n}{E^{t-1}}) \frac{q}{N-1}, \quad \text{and} \tag{3}$$

$$P({k_i}^t = n \mid {k_i}^{t-1} = n+1) = p_0 \frac{n+1}{E^{t-1}} + q \frac{n+1}{E^{t-1}}. \tag{4}$$

We are now able to write down the difference equation describing the expected probability distribution for the copy number of the i^{th} mer:

$$\begin{aligned}
P({k_i}^t = n) &= P({k_i}^{t-1} = n-1) P({k_i}^t = n \mid {k_i}^{t-1} = n-1) \\
&\quad + P({k_i}^{t-1} = n) P({k_i}^t = n \mid {k_i}^{t-1} = n) \\
&\quad + P({k_i}^{t-1} = n+1) P({k_i}^t = n \mid {k_i}^{t-1} = n+1) \\
&= P({k_i}^{t-1} = n-1) \left(p_1 \frac{n-1}{E^{t-1}} + (1 - \frac{n-1}{E^{t-1}}) \frac{q}{N-1} \right) \\
&\quad + P({k_i}^{t-1} = n) \left(1 - \frac{n}{E^{t-1}} - (1 - \frac{n}{E^{t-1}}) \frac{q}{N-1} \right) \\
&\quad + P({k_i}^{t-1} = n+1) \left(p_0 \frac{n+1}{E^{t-1}} + q \frac{n+1}{E^{t-1}} \right) \\
&= P({k_i}^{t-1} = n-1) \left((p_1 - \frac{q}{N-1}) \frac{n-1}{E^{t-1}} + \frac{q}{N-1} \right) \\
&\quad + P({k_i}^{t-1} = n) \left(1 - (1 - \frac{q}{N-1}) \frac{n}{E^{t-1}} - \frac{q}{N-1} \right) \\
&\quad + P({k_i}^{t-1} = n+1) \left(p_0 \frac{n+1}{E^{t-1}} + q \frac{n+1}{E^{t-1}} \right).
\end{aligned} \tag{5}$$

Since the total number of mers in a genome is usually very large, and each mer species only accounts for a very small fraction of the genome, we assume that the copy number of each mer species evolves independently. Therefore, the above equation can be viewed as an expression of the copy number distribution of all possible mers in a genome. This assumption is validated by Monte Carlo simulations.

[4] We approximate the probability of a specific mer being chosen to substitute another mer during substitution as $\frac{1}{N-1}$ (instead of $\frac{1}{3l}$). This approximation stands when mer size l is small.

4.2 Model fitting

The simple graph model is applied to fit the mer-frequency distributions from various genomes. Since the process is non-stationary, we use numerical simulation in the model fitting. The initial condition is set as a proto-genome: a random sequence of size 1Kb containing uniformly distributed mers. The iteration proceeds until the final genome size reaches the real genome size under study. Some of the fitted mer-frequency distribution results can be seen in Figure 1. The model has two degrees of freedom in its parameter space. Here we choose to estimate q and $\frac{p_1}{p_0}$. The parameters for the optimal fit to the data from genome analysis are estimated so that the sum of the absolute differences between the real data and the data produced by the model are minimized. Some of the fitted parameter values are shown in 1.

Mer size	**6-mer**		**7-mer**		**8-mer**		**9-mer**	
	q	p_1/p_0	q	p_1/p_0	q	p_1/p_0	q	p_1/p_0
M. genitalium	0.0176	1.1	0.0436	1.1	0.1587	1.5	0.3222	1.5
M. pneumoniae	0.0319	1.1	0.1151	1.5	0.2309	1.5	0.4363	1.5
P. abyssi	0.0269	1.1	0.0672	1.2	0.1778	1.4	0.3897	1.5
P. horikoshii	0.0234	1.1	0.0443	1.1	0.139	1.3	0.3456	1.5
P. furiosus	0.0213	1.1	0.0384	1.1	0.1119	1.2	0.3114	1.5
H. pylori	0.018	1.1	0.032	1.1	0.0925	1.3	0.2262	1.5
H. influenzae	0.0202	1.1	0.0366	1.1	0.1364	1.5	0.2802	1.5
S. tokodaii	0.018	1.1	0.032	1.1	0.0925	1.3	0.2262	1.5
S. subtilis	0.0187	1.1	0.0326	1.1	0.1139	1.4	0.2585	1.5
E. coli K12	0.0207	1.1	0.0334	1.1	0.0698	1.1	0.2389	1.5
S. cerevisiae	0.0113	1.1	0.0176	1.1	0.0459	1.2	0.1311	1.4
C. elegans	–	–	0.0076	1.1	0.0115	1.1	0.0275	1.2

Table 1. Graph model parameters (q, p_1/p_0) fitted to the mer-frequency distribution data (6 to 9-mer) from the whole genome analysis.

The fitted parameters in the table show some interesting properties. The optimal relative substitution probabilities, q values, of a particular genome increase monotonically with the mer-size (l). This may reflect the scaling effect in this simple model introduced by fixing the size of duplication or deletion as the size of one mer. In the related biological processes, while one substitution changes one mer to another, one duplication or deletion may change the copy numbers of more than one mer. In a particular duplication or deletion event, when the mer-size increases, the corresponding number of mers being affected by the process decreases. Therefore, the relative probabilities of substitution of larger mer's tend to be larger than those of the smaller mer's. It is also noticeable that the values of the parameter $\frac{p_1}{p_0}$ increase along with the mer-sizes in each genome. This suggests that duplication probability p_1 decays

more slowly than deletion probability p_0 when the mer size increases. Such a behavior indicates that duplications of large regions occur more often than deletions of large regions. Furthermore, the ratios $\frac{p_1}{p_0}$ are consistently larger than 1, which validates our assumption of $p_1 > p_0$ in the model.

The organisms listed in the table 1 are ordered by their genome sizes (from 580Kb for *M. genitalium* to 97Mb for *C. elegans*). There is a slight tendency for q values to be smaller when the genome sizes become larger. This may reflect the fact that the sizes of duplication or deletion units increase with the corresponding genome size. However, there are exceptions: For example, *M. genitalium* has a genome of size 580Kb, its q values are smaller than many listed organisms with larger genome sizes, such as *M. pneumoniae* (816Kb), *P. abyssi* (1.8Mb), *etc*. This observation may be explained by a higher substitution rate in *M. genitalium*.

Under biologically reasonable assumptions, the parameters of the model can be used to estimate the size distribution of the duplication and deletion events in real genomes. Previous research [24] [25] has shown that the size distribution of the insertion and deletion regions in the genomes examined follows a power-law. The exponents of the power-law are the key characterizing factors of the size distributions, and are of the greatest interest as they reveal the link between the genome dynamics and genome statistics. Unfortunately, direct estimation of the exponents (e.g., from sequence comparison) not only requires complex and expensive computation, but also imposes strong constraints on data sources so as to minimize ambiguity. However, for a specific genome, our model and the mer distribution data are sufficient to determine the exponents reasonably well, as described below. We start with the following assumptions: During genome evolution, the averaged rate of point mutation in each time interval is μ per nucleotide; the probability of duplicating a fragment of size x in each time interval is $f_1 x^{-b_1}$; the probability of deleting a fragment of size x in each time interval is $f_0 x^{-b_0}$ (Here, f_1 and f_0 are normalization constants; b_1 and b_0 are the exponents for the power-law distributions of duplication and deletion sizes, respectively.). The relationships between the model parameters and the size of the mer (l) they are fitted for in a specific genome are shown in equations (6), (7), and (8).

When the model and its parameters are fitted to a specific genome for a sufficiently large number of mer-sizes, the exponents (b_1 and b_0) can be estimated by a linear regression using the relationship between the parameter ratios and mer sizes (l) from equations (6), (7), and (8). This approach allows us to infer the size distribution of duplication and deletion events over the evolutionary history of that genome.

$$\frac{q}{p_1} = \frac{P(\text{an } l \text{ mer gets substituted})}{P(\text{an } l \text{ mer gets duplicated})} = \frac{\mu l}{\int_{x \geqslant l}^{\infty} f_1 x^{-b_1}(x/l)dx}$$
$$= \frac{\mu l^2}{\int_{l}^{\infty} f_1 x^{1-b_1} dx}$$
$$= \frac{\mu(b_1-2)}{f_1} l^{b_1} \propto l^{b_1} \tag{6}$$
$$\frac{q}{p_0} = \frac{P(\text{an } l \text{ mer gets substituted})}{P(\text{an } l \text{ mer gets deleted})} = \frac{\mu l}{\int_{x \geqslant l}^{\infty} f_0 x^{-b_0}(x/l)dx}$$
$$= \frac{\mu(b_0-2)}{f_0} l^{b_0} \propto l^{b_0} \tag{7}$$
$$\frac{p_1}{p_0} = \frac{P(\text{an } l \text{ mer gets duplicated})}{P(\text{an } l \text{ mer gets deleted})} = \frac{\int_{x \geqslant l}^{\infty} f_1 x^{-b_1}(x/l)dx}{\int_{x \geqslant l}^{\infty} f_0 x^{-b_0}(x/l)dx}$$
$$= \frac{f_1(b_0-2)}{f_0(b_1-2)} l^{b_0-b_1} \propto l^{b_0-b_1} \tag{8}$$

4.3 Polya's model

Although the parsimonious model described above captures the most important elements during genome evolution, it omits most of the details. To get a more comprehensive and specific understanding of genome evolution, we develop a more realistic model. The model will mainly include the parsimonious rules, but apply them in a more interactive way. The model is an extension of the Polya's urn model on a string. In this model, the same three main events in evolution are considered: *duplication*, *deletion*, and *substitution* (Figure 6). Similar to the simple model, genome evolution is modeled as a stochastic process that goes through multiple iterations. Within each iteration, one of the three events happens with a certain probability. However, unlike in the simple model, the details of the events can also be manipulated (Figure 6). In every iteration, a set of probability distributions are applied to decide the changes in the *in silico* evolution. All the probability distribution functions can be inter-dependent, as well as independent.

As the model approaches in its resemblance to reality, it becomes increasingly complex, thereby making the explicit mathematical approach infeasible. Therefore, large-scale *in silico* experiments are needed. Finally, genome evolution is also a population process. To understand the genome evolution more completely, we also plan to simulate it in a population model which integrates natural selection and polymorphism effects.

5 Conclusion

Among the few fundamental "dogmas" at the core of biological sciences, a central and most elegant one is likely to be the "evolution by duplication."

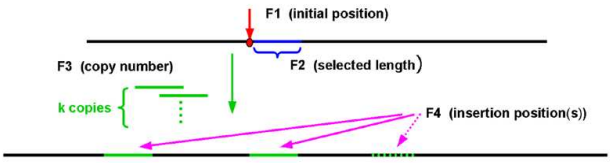

Fig. 6. The more realistic model in our *in silico* simulation. During each iteration in the simulation, just like in the simple graph model, one of the three events (deletion, duplication, or substitution) happens with probabilities p_0, p_1 and q. A probability distribution (F_1) decides the initial position in the genome where a chosen event will happen. Another probability distribution function (F_2) controls the size of the fragment chosen from the existing genome if either a duplication or a deletion event happens. A third probability distribution function (F_3) decides the copy number for the duplication. (When the copy number is one, a translocation happens.) And a fourth probability distribution function (F_4) decides the insertion site(s) for the duplicated fragment(s) during a duplication event. The distribution functions can be interdependent. The model is a realization of Polya's Urn on a string.

For many years, this theory is likely to remain intriguing and mysterious in its pervasive power in explaining many seemingly unrelated biological phenomena. Our efforts to understand it better will continue to raise many beautiful mathematical and computational questions requiring many novel techniques.

For several years, we have focused on these problems, and have developed many computational techniques, not discussed in the paper but briefly mentioned below: *Valis*, a computational environment and language allowing us to rapidly prototype genome-analysis algorithms and visualization tools; *Genome Grammar*, a highly memory and run-time efficient tool for large scale *in silico* evolution simulations; *Simpathica*, a tool for understanding biological processes involved in genome evolution and their effects on pathways.

References

1. Peng, C.K. et al: Long-range correlations in nucleotide sequences. Nature **356**, 168–170 (1992)
2. Gomez, S.M., Rzhetsky, A.: Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome. Bioinformatics **17**, 988–996 (2001)
3. Fields, S., Schwikowski, B., Uetz, P.: A Network of protein-Protein interactions in Yeast. Nature Biotechnol. **18**, 1257-1261 (2000)
4. Albert, R., Barabasi A-L.: Statistical mechanics of complex networks. Review of Modern Physics **74**, 48–97 (2002)
5. Havlin, S. et al: Mosaic organization of DNA nucleotides. Physical Review E. **49**, 1685–1689 (1994)
6. Ehrlich, S.D., Viguera, E., Canceill, D.: Replication slippage involves DNA polymerase pausing and dissociation. EMBO J. **20**, 2587–2596 (2001)

7. Lilley, D.M.J.,Eckstein, F.: *DNA Repair* (Springer, Berlin Heidelberg new York 1998)
8. Albert, R. et al: The large-scale organization of metabolic networks. Nature **407**, 651–654 (2000)
9. Barabasi, A.L. et al: Lethality and centrality in protein networks. Nature **411**, 41–42 (2001)
10. Gerstein, M., Qian, J., Luscombe, N.M.: Protein family and fold occurrence in genomes: power-law behavior and evolutionary model. Journal of Molecular Biology **313**, 673–681 (2001)
11. Rain, J.C. et al: The protein-protein interaction map of *Helicobacter pylori*. Nature **409**, 211–215 (2001)
12. Vogelstein, B., Lane, D., Levine, A.J.: Surfing the P53 network. Nature **408**, 307–310 (2000)
13. Johnson, N.L.: *Urn models and their Application* (Wiley 1977)
14. Ganapathiraju, M. et al: Comparative n-gram analysis of whole-genome protein sequences. In: *HLT'02: Human Language Technologies Conference,* San Diego, California, USA, March 2002.
15. Ohno, S.: *Evolution by Gene Duplication* (Springer, Berlin Heidelberg New York 1970)
16. Apweiler, R. et al: The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Research **29**, 37–40 (2000)
17. Sole, R.V., Pastor-Satorra, R., Smight, E.: Evolving protein interaction networks through gene duplication. Santa Fe Institute Working Paper 02-02-008 (2002)
18. Mantegna, R.N. et al: Linguistic features of noncoding DNA sequences. Physical Review Letters **73**, 3169–3172 (1994)
19. Sneppen, K., Maslov, S.: Specificity and stability in topology of protein networks. Science **296**, 910–913 (2002)
20. Buldyrev, S.V. et al: Fractal landscapes and molecular evolution: modeling the myosin heavy chain gene family. Biophysical Journal **65**, 2673–2679 (1993)
21. Eichler, E.E.: Recent duplication, domain accretion and the dynamic mutation of the Human genome. Trends in Genetics **17**, 661–669 (2001)
22. Bailey, J.A. et al: Recent segmental duplications in the Human genome. Science **297**, 1003–1007 (2002)
23. Graur, D., Li, W-H.: *Fundamentals of Molecular Evolution* (Sinauer 2000)
24. Gu, X., Li, W-H.: The size distribution of insertions and deletions in Human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment. Journal of Molecular Evolution **40**, 464–473 (1995)
25. Ophir, R., Graur, D.: Patterns and rates of indel evolution in processed pseudogenes from Humans and Murids. Gene **205**, 191–202 (1997)