Models of Genome Evolution

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

[1] Biology Department, New York University.
[2] Watson School of Biological Sciences, Cold Spring Harbor Laboratory.
[3] Courant Institute, New York University,
251 Mercer Street,
New York, NY 10012, USA.
{joey,mishra}@cs.nyu.edu,
WWW home page:http://www.cs.nyu.edu/cs/faculty/mishra/

Abstract. In 1970, Susumu Ohno proposed an evolutionary theory, called "evolution by duplication" that can now be verified with the available genome sequences.

Recently, several mathematical models have been proposed to explain the topology of protein interaction networks. The power-law connectivity 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 models the new proteins emerging from the older ones, with a propensity to interact with only those proteins that its other family members interact with. Since the protein interaction network, as well as other higher-level cellular processes, is encoded in genomic sequences, the evolutionary structure, topology and statistics of many biological objects (pathways, phylogeny, symbiotic relations, etc.) are rooted in the genomic sequence's evolution dynamics.

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 number of the inserted fragments and the sequence 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 (Deoxyribo Nucleic Acid), and forms the genetic core of a cell. The genes ultimately encode for the protein (a chain of amino acids) and in turn, the genes are regulated by transcription factors and other operons, many of which are proteins.

The sequence 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* (pair of proteins), *trimers* (triplets) and *multimers*. An *isoform* of a protein is a slightly different protein with closely related sequence, and often share similar functional properties, e.g., enzymatic reactions, but are regulated differently. Nevertheless, 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.

However, 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, etc.—constantly act on genomes and drive the genomes to evolve dynamically. These alternations in the genomic sequences can further lead to the corresponding changes in the higher-level cellular information (transcriptome, proteome, interactome), and are crucial in explaining the myriads 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 development of a rigorous, faithful, and yet simple abstract model for genome evolution.

All the present genomes can be viewed as a snapshot of an ongoing genome evolution process. Although 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, currently available. With the development of high throughput experimental technology, the flow of information at different levels of biology (genome, proteome, transcriptome, interactome) is increasing dramatically. By analyzing and comparing those data, we are now able to look for the structure of cellular information compilation and the dynamics of evolution 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 on 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], 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 featured 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 positively correlated 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 reflections of an underlying generic structure of life in different aspects. See Genome Analysis section for more examples. Given that different organisms can experience drastically different evolutionary environments, yet a common structure in their cellular processes still preserves, such a structure may represent one of the most basic rules 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 and a more realistic, thus complicated version of the model in large-scale in silico evolution simulation. The simple model fits nicely with real-world data for mer-frequency distributions. These results suggest that despite the highly diversified evolutionary environment for different organisms, the essential composition of the evolutionary dynamics is commonly shared. The 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 results of a simple stochastic evolutionary process.

2 Evolution by duplication

Susumu Ohno proposed an evolutionary theory "evolution by duplication" [15]. Although not clearly noted, his theory proposed a "rich gets richer" rule in genome evolution. The theory argues that evolutionary advantage of evolution by duplication is 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 to obtain a new gene that is advantageous. Gene duplication can speed up the search for higher fitness in various ways [23]: It can adjust gene dosage; or attain a permanent heterozygous advantage by incorporation of two former alleles into the genome; or allow more specialized function by differential regulation of the duplicated genes; or create a new gene with 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 duplications 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 DNA polymerase to pause, dissociate and continue an erroneous strand extension after reassociation. During meiotic cell devisions, recombinations happen between two DNA molecules through crossing-overs between corresponding homologous regions. However, unequal crossing-over between two DNA molecules bearing successive repeated fragments will result in duplication or deletion of the repeated units in the daughter cells [23]. Another process that can introduce duplications or deletions of relatively large size and globally in the genome is by mobile DNA elements (insertion elements, transposons, and retrotransposons)s [23]. The mobile elements can be either excised or copied out of its original position, and insert themselves or their copies elsewhere in the genome where target sequences can be found. The frequencies and sizes of the deletions or insertions vary correspondingly to specific elements. Since the target sequences are widely distributed in the whole genome, the mobile elements can essentially affect the sequence changes on the whole genome range. Therefore, 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 has DNA repair machineries to counteract the changes made in the sequences and prevents genomes from changing too rapidly [7]. For example, mismatch repair (MMR) is mainly responsible for correcting most of the deletions and insertions of various sizes. Therefore, duplications, deletions, and similarly other changes in genomes are the results of the counteractions 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 the 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-enrichment of some elements.

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" [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 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 the protein interaction network, as well as other higher-level cellular information, is encoded in genomic sequences, the evolution of its 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.

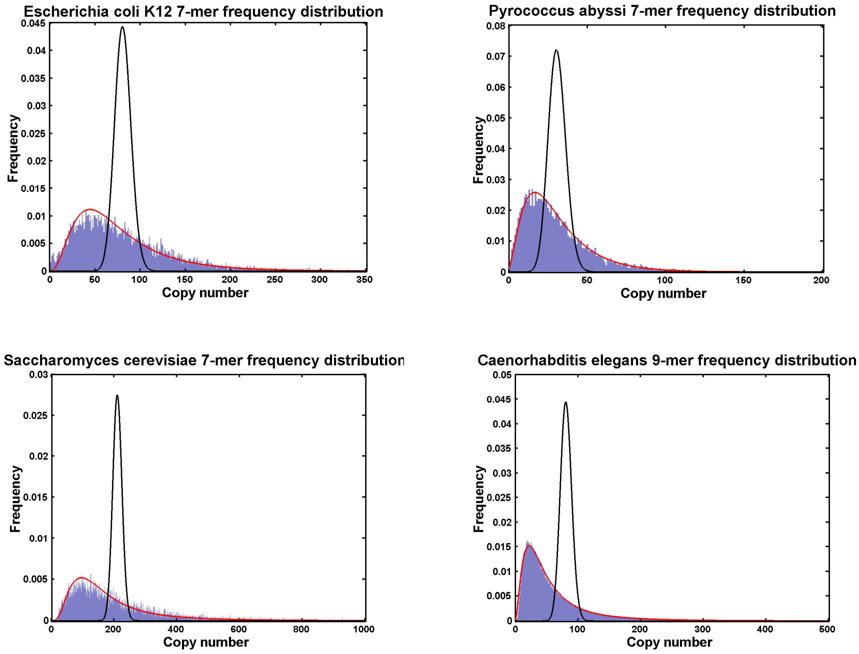

Fig. 1. Mer-frequency distribution in some genomes examined. The plots shows the non-overlapping mer frequency distribution (blue bars) in the genomes of a eubacteria (*E.coli* K12), an archea (*P.abyssi*) and two eukaryote (*S.cerevisiae* and *C.elegans*). When compared with the expected distribution 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.

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 evidences 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 on the distribution of the end-points of putative large segmental duplications in human genome provides convincing evidences that duplications tend to occur more often around the previous 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, meanwhile the other segments are likely to be more suppressed. The duplication dynamic implied by end-points distribution analysis may explain our observation in the distribution of protein family sizes, following a power-law distribution and 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 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 a 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 random distribution (see Figure 1 for some of the results). Furthermore, they are characterized by the same type of deviation—over-representation of high frequency mers.

We also have looked at the mer-frequency distribution in coding sequences only, 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 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 mer's in genomes and short amino-acid words in proteomes can be simply due to the duplications processes during genome evolution.

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

As mentioned in the Introduction section as well as our mer-frequency analysis, the results of various statistical analyses on genomic data have suggested that there is a generic evolution dynamic dominated by duplications. To justify the hypothesis, it is important to study the dynamic of duplication processes itself. Although the exact molecular mechanisms causing duplications are not fully understood, we can approach the problem indirectly by looking at the distribution of the most recent segmental duplications in genomes.

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

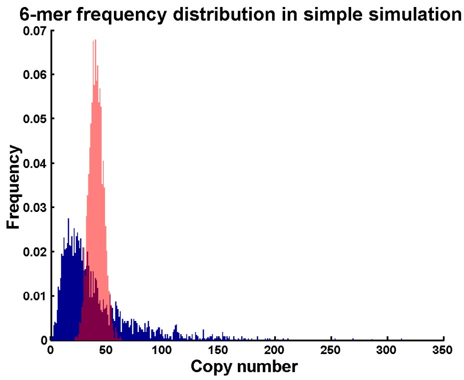

Fig. 2. The 6-mer frequency distribution of the resulting sequence of a simple "evolution by duplication" simulation. The initial condition is a random sequence of length 1000bp. The sequence is evolved through multiple iterations until it reaches a length of 500Kb. In each iteration, a fragment is randomly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

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

Since the end-points can be viewed as the signatures left by the duplication processes over the genome evolutionary history, their distribution along the genome could reveal some dynamic features of 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" on the corresponding genomic region. When we plotted the histogram of the end-point density from a chromosome or from the whole genome, we discovered another power-law distribution (Figure 3). It can be interpreted as that duplications tend to happen more often at the previous duplication sites, driven by a positive feedback dynamics. 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/on some previous duplicated segments, and are absent elsewhere.

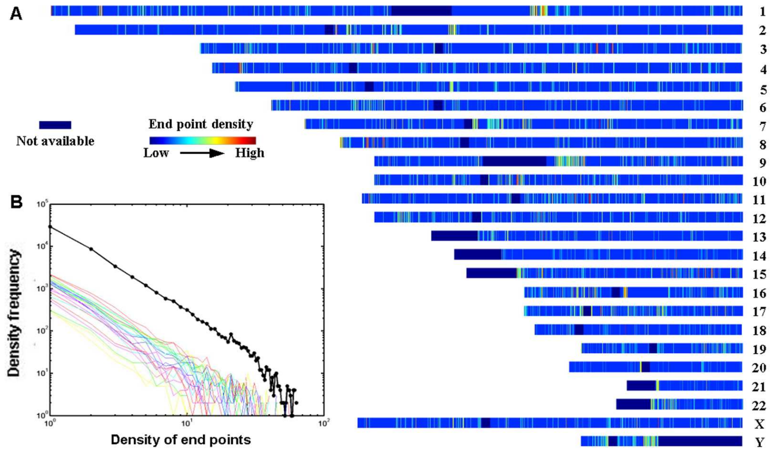

Fig. 3. The distribution of the potential duplication 'hot-spots' on human genome. A. The distribution of the duplication segment end-points on the chromosomes (here shows window size of 1Kb). The 'hot-spots' density is color coded (see the color bar). The dark areas represent the chromosome regions 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 a non-overlapping window size of 1Kb (density of end-points), starting from 1. The Y axis indicates the number of non-overlapping 1Kb windows containing a certain number of end-points (density frequency). It is clear that the Log10(density of end-points) and Log10(density frequency) form a linear relationship, both in the whole-genome wide range (thick black points) and in individual chromosome range (various thin colorful lines).

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 verify whether the generic structure observed in such 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 proteins or domains are mostly above 50 amino acids (150 nucleotides), which are on a bigger scale than that of the mer analysis. The analysis of protein/domain families is based on sequence signature and homology. A family of protein/domain found by this method can be viewed as a cluster of amino acid sequences from proteome that share enough similarity with each other and have maintained their critical sequences. When the histograms of the sizes of those protein 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*, *etc.* (Figure 4). Therefore, the protein family size distribution, or more generally, the cluster size of homologous amino acid sequences seems to follow a power-law distribution.

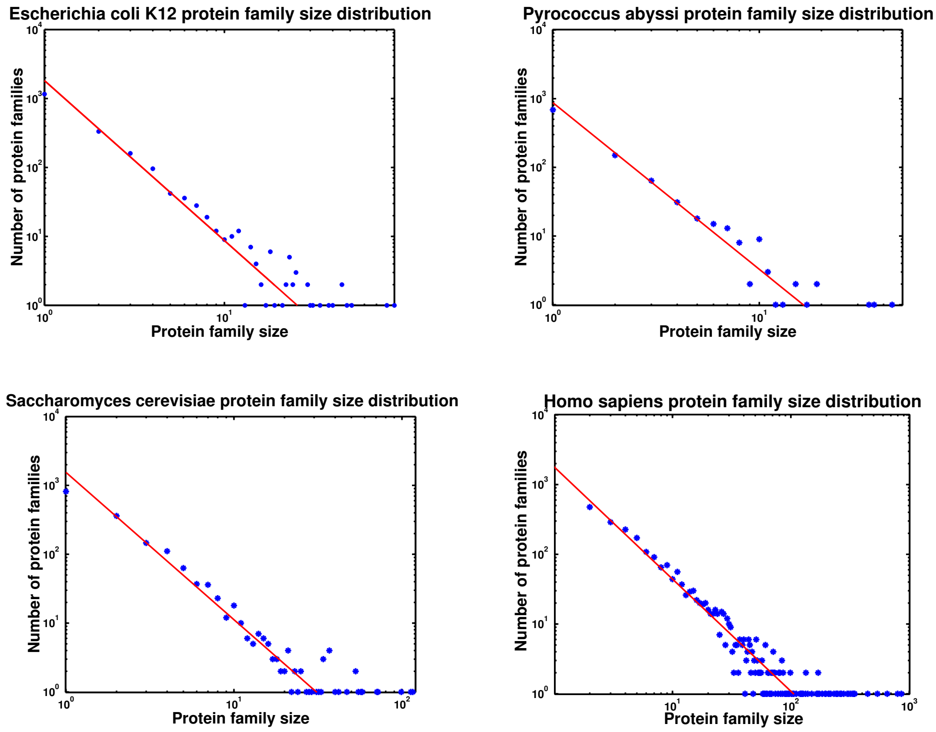

Fig. 4. Protein family size distribution in some genomes examined. The plots show the protein family size distribution in the corresponding proteomes (blue dots) . The protein family data is extracted from Interpro database. The plots are on a log-log scale. The almost linear shape (red line) on such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

Here we observed the same deviation pattern from random distribution on this larger scale as in the small-scale mer distribution analysis—further confirming 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 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 omission of any of the three processes renders the model unsuitable to fit the 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 mer 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 5' end to 3' end. Therefore, the number of directed edges from i^{th} node to j^{th} node ($k_{i,j}$) 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 i^{th} node, since due to Eulerian property of the 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 totally $N = \frac{4^l}{2}$ 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 mers are randomly distributed. Thus, the initial graph is a random graph with 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 $p_1 > p_0$. The biological processes that can cause duplications or deletions include homologous or heterologous recombination or 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 such assumption, 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

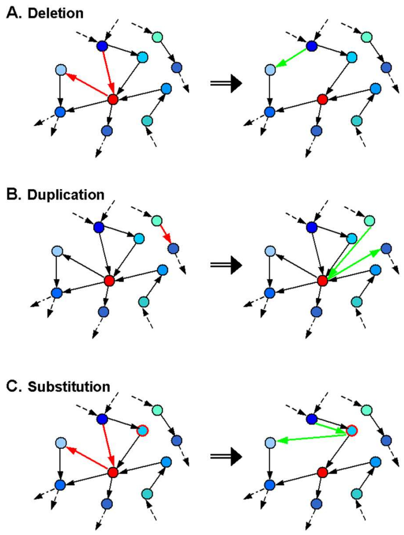

Fig. 5. The three processes during graph evolution: *deletion*, *duplication* and *substitution*. In each process, the target node (indicated by red fill in the figure) is chosen with preference for nodes with larger degrees: If the *i*-th node has degree k_i, the probability of it being chosen is proportional 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.

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

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

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

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

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

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

Since the total number of mers in a genome is usually very large, and the total number of species of mers of a particular length is also very large when the mer length is big enough, we can assume that the probability distributions of the copy number of each species of mers are independent. Thus, the expected probability distribution of the copy number of a particular mer can represent the expected mer copy number distribution of the genomes.

4.2 Model fitting

The simple graph model is applied to fit the mer-frequency distribution 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 mer's. 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 such 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 the following table 1.

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

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. Fitted parameters (q, p_1/p_0) in the graph model to the mer-frequency distribution data (6 to 9-mer) from the whole genome analysis.

or deletion may change the copy numbers of more than one mer's. In a particular duplication or deletion event, when the mer-size increases, the corresponding number of mer's being affected by the process decreases. Therefore, the relative probabilities of substitution of larger mer's tend to be larger than those of of the smaller mer's. Also noticeable is that the values of the parameter $\frac{p_1}{p_0}$ increase along with the mer-sizes in each genome. Such a behavior indicates that the sizes of duplication events mostly are bigger than deletions, 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 *M.genitalium* to 97Mb *C.elegans*). There is a slight tendency for q values to be smaller when the genome sizes become larger. This may reflect that the sizes of duplication or deletion 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*.

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, pre-decided probability distributions are applied to realize the changes in the in silico evolution. All the probability distribution functions can be inter-dependent, as well as independent.

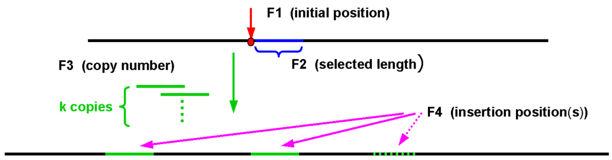

Fig. 6. The more realistic model in our in silico simulation. During each iteration in the simulation, just like in the simple graph model, one of the three events (deletion, duplication, or substitution) happens with probabilities p_0, p_1 and q. A probability distribution (F_1) decides the initial position in the genome where a chosen event will happen. Another probability distribution function (F_2) controls the length 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.

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." 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., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Sciortino, F., Simon, M., and Stanley, H.E.: Long-range Correlations in Nucleotide Sequences. Nature (1992), **356**: 168–170.
2. Gomez, S.M., Rzhetsky, A.: Birth of Scale-free Molecular Networks and the Number of Distinct DNA and Protein Domains per Genome. Bioinformatics (2001), **17**: 988–996.
3. Fields, S., Schwikowski, B., Uetz, P.: A Network of Protein-Protein Interactions in Yeast. Nature Biotechnol. (2000), **18**: 1257-1261.
4. Albert, R., Barabasi A-L.: Statistical Mechanics of Complex Networks. Review of Modern Physics (2002), **74**: 48–97.
5. Havlin, S., Simons, M., Stanley, H.E., Goldberger, A.L., Peng, C.K., Buldyrev, S.V.: Mosaic Organization of DNA Nucleotides. Physical Review E. (1994), **49**: 1685–1689.
6. Ehrlich, S.D., Viguera, E., Canceill, D.: Replication Slippage Involves DNA Polymerase Pausing and Dissociation. EMBO J. (2001), **20**: 2587–2596.
7. Lilley, D.M.J.,Eckstein, F.: **DNA Repair**. Springer-Verlag, 1998.
8. Albert, R., Oltvai, A.N., Barabasi, A.L., Jenogh, H., Tombor, B.: The Large-scale Organization of Metabolic Networks. Nature (2000), **407**: 651–654.
9. Barabasi, A.L., Oltavai, Z.N., Jeong, H., Mason, S.P.: Lethality and Centrality in Protein Networks. Nature (2001), **411**: 41–42.
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 (2001), **313**: 673–681.
11. Rain, J.C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schachter, V., Chemama, Y., Labigne, A., Legrain, P.: The Protein-Protein Interaction Map of *Helicobacter pylori*. Nature (2001), **409**: 211–215.
12. Vogelstein, B., Lane, D., Levine, A.J.: Surfing the P53 Network. Nature (2000), **408**: 307–310.
13. Johnson, N.L.: **Urn models and their Application**. John Wiley & Sons, 1977.
14. Ganapathiraju, M., Weisser, D.: Comparative n-gram Analysis of Whole-genome Protein Sequences. Personal communication, (2002).
15. Ohno, S.: **Evolution by Gene Duplication**. Springer-Verlag, 1970.
16. Apweiler, R., Attwood, T.K., *et al.*: The InterPro Database, an Integrated Documentation Resource for Protein Families, Domains and Functional Sites. Nucleic Acids Research (2000), **29**: 37–40.
17. Sole, R.V., Pastor-Satorra, R., Smight, E.: Evolving Protein Interaction Networks through Gene Duplication. Santa Fe working paper (2002).
18. Mantegna, R.N., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.K., Simons, M., Stanley, H.E.: Linguistic Features of Noncoding DNA Sequences. Physical Review Letters (1994), **73**: 3169–3172, 1994.
19. Sneppen, K., Maslov, S.: Specificity and Stability in Topology of Protein Networks. Science (2002), **296**: 910–913.
20. Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.K., Stanley, H.E., Stanley, M.H., Simons, M.: Fractal Landscapes and Molecular Evolution: Modeling the Myosin Heavy Chain Gene Family. Biophysical Journal (1993), **65**: 2673–2679.
21. Eichler, E.E.: Recent Duplication, Domain Accretion and the Dynamic Mutation of the Human Genome. Trends in Genetics (2001), **17**: 661–669.
22. Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D., Myers, E.W., Li, P.W., Eichler, E.E.: Recent Segmental Duplications in the Human Genome. Science (2002), **297**: 1003–1007.
23. Graur, D., Li, W-H.: **Fundamentals of Molecular Evolution**, Sinauer Associates, Inc. (2000).