EVOLUTION WITHOUT SEX, DRUGS AND BOOLEAN FUNCTIONS

NISHEETH K. VISHNOI

Enough cannot be said about Evolution, a theory for which was first formulated by Darwin and Wallace, which drives life and, perhaps, is the reason for the emergence of the genetic information that forms its basis. At a very high level, evolution is an iterative process which takes feedback from an environment in order to refine existing information and bias it towards that which is more likely to be replicated. However, the evolutionary forces that have been active on Earth for around four billion years have learned not to put all eggs in one basket; maintaining *diversity* at the expense of *purity* is important in order to make information more robust against sudden environmental changes. Thus, critical to evolution is a replication mechanism that introduces variability.

In its most basic form, the information about life on Earth is organized in molecules such as RNA and DNA. An RNA molecule is a sequence composed of four nucleotides A, G, C and U while a DNA molecule consists of a pair of complementary sequences over the alphabet A, G, C and T. The replication of RNA molecules is primitive and error-prone when compared to that of DNA. DNA replicators have the ability to check and correct errors while copying, something RNA replicators cannot do, making DNA replication a more robust and stable process. Thus, from a diversity point of view, RNA seems to be at an advantage. DNA, however, has found a sophisticated strategy to introduce diversity, namely, *sex* or *recombination. Horizontal gene transfer*, responsible for bacterial antibiotic resistance, is another mechanism to introduce diversity. I will focus on abstinent, RNA-type evolution; more can be found on sex, its advantages and how CS and Math can contribute in ongoing work at Berkeley [LPDF08, CLPV12].

How can an error-prone copying mechanism be advantageous? After all, wouldn't errors corrupt the genetic information in RNA? To understand this, notice that sequences differ in their *fitness*, i.e., their ability to survive and replicate in a given environment. Hence, the above would indeed be true if a genetic sequence was the most fit in its current environment. However, it is unlikely that, at the genesis of information, the optimal sequence was simply *handed over*. It is more likely that life started with sub-optimal genetic information and erroneous copying gave the ability to explore the more-than-astronomically-sized space of all sequences of A, G, C and U in a random manner. In this process, RNA can discover sequences that are more likely to replicate; the imperfect copying mechanism and feedback via fitness gives rise to improved genetic information through successive generations. An example of this phenomena occurred in Mexico around 1545 where it is believed that a strain of Cocoliztli, an RNA virus, mutated and gained the ability to be transmitted from humans to humans instead of rats to humans. [1] As a result, the virus wiped out up to 70% of the defenseless Aztec population in a matter of few years.

Abstraction and modeling form the basis of any scientific inquiry as they allows one to ask if-then questions and conduct thought and, these days, computer experiments. Many times, a good model or theory has the ability to change lives. Manfred Eigen, a chemist, did exactly this in his seminal papers from the '70's [Eig71, ES77] which address evolution of genetic information as described above. He introduced the *quasispecies model*, a remarkably clean mathematical model that captured the error-prone evolution of molecules such as RNA and addressed the origins-of-life question. It was argued, see [Eig93], that an important property in such a model of evolution is the existence of the *error threshold*, i.e., a replication error rate below which genetic information is intact and above which it is destroyed. Besides the insights the quasispecies model provides on the emergence of life, it has had a powerful impact on design of drug strategies for RNA-viruses such as Foot-and-Mouth-Disease [HD98], Poliovirus [CCA01] and even RNA *retroviruses* such as HIV [LEK$^+$99]. In all these cases, the error threshold phenomena is leveraged to attempt to mutate the virus to death.

[1] Typically, an RNA-virus lacks the ability to copy itself. When a virus has the *right* proteins to enter a living cell, it does so and hijacks its copying machinery, disrupting the normal function of the cell and, hence, the life-form the cell is a part of. The proteins that form the virus are also encoded in its RNA. Proteins themselves are nothing but sequences of amino acids and, rather remarkably, this sequence completely determines how it folds. The protein, once folded, becomes like a nano-machine and performs all kinds of mechanical functions. A small error in the part of virus RNA that encodes a particular protein results in a slightly modified amino acid sequence which means the protein folds a bit differently. Hence, its mechanical behavior changes and it can gain abilities it previously did not possess, in this case, the ability to hijack human cells directly.

The rest of this overview is organized as follows: First, I will describe the quasispecies model and the associated phenomenon of error threshold. Subsequently, I will discuss a stochastic, finite population model for evolution which incorporates observed stochastic effects. In the process, I will point out several important mathematical and computational problems whose resolution is currently assumed without due rigor. The problems involve topics such as Boolean functions, threshold phenomena, and random walks and, should be of general interest to the TCS and Math community. I sincerely hope that many of these will be resolved and, with more ideas coming in, we can contribute to an area of science which holds the potential to explain why we are and help us continue to be.

The Quasispecies Model and the Error Threshold. Let us consider the quasispecies model for the evolution of strings of length L composed of two building blocks 0 and 1 which gives rise to the sequence space $\{0,1\}^L$. The environment remains fixed throughout this evolutionary process and the fitness of sequences in this environment is described by a function a from $\{0,1\}^L$ to the positive reals, often referred to as the *fitness landscape*. In other words, the sequence space is the *Boolean hypercube* and the fitness landscape a *Boolean function*. The model assumes that, while copying a sequence, each of its bits is flipped with probability μ which is called the *error* or the *mutation rate*. Hence, the probability that a string σ mutates to τ is $Q_{\sigma\tau} := \mu^{d_H(\sigma,\tau)}(1-\mu)^{L-d_H(\sigma,\tau)}$ where $d_H(\sigma,\tau)$ is the Hamming distance between σ and τ, i.e., the number of positions where σ and τ differ, implying that $Q_{\sigma\tau} = Q_{\tau\sigma}$. The model assumes infinite population and, hence, at any point keeps track of only the fraction of sequences of each type. Evolution starts with an arbitrary distribution over $\{0,1\}^L$ and it iterates applying the tenets of reproduction-selection-mutation as follows: In the *reproduction* stage each sequence σ in the current population produces a_σ copies of itself. Since the total mass could become more than one, in the *selection* stage we normalize the mass of each sequence to bring the total mass back to one unit. The *mutation* step can be thought of as deterministic since the population is infinite: One unit of population of σ gives rise to $Q_{\sigma\tau}$ fraction of population of τ. Thus, the eventual outcome of the evolutionary process is not a single sequence (i.e., a species), rather it is an invariant distribution over sequences, justifying the term *quasispecies*.

More formally, suppose $\mathbf{m}^0$ is the starting distribution over the population at time 0 and m^t_σ is the fraction of the population of σ at time $t \geq 0$. Combining the reproduction and mutation step, for every string τ, $a_\tau m^t_\tau Q_{\tau\sigma}$ (fractional) copies of σ are produced (without normalization). Thus, the population evolves via the following coupled difference equations: $\forall \sigma \in \{0,1\}^L$, $m^{t+1}_\sigma := \frac{\sum_{\tau\in\{0,1\}^L} m^t_\tau a_\tau Q_{\tau\sigma}}{\sum_{\tau\in\{0,1\}^L} m^t_\tau a_\tau}$. Let Q be the matrix with rows and columns indexed by sequences in $\{0,1\}^L$ where the entry corresponding to row σ and column τ is $Q_{\sigma\tau}$ as above. Further, let A be the diagonal matrix with $A_{\sigma\sigma} := a_\sigma$. Then the evolutionary equations above can be succinctly expressed in the matrix form as $\mathbf{m}^{t+1} := \frac{QA\mathbf{m}^t}{\|A\mathbf{m}^t\|_1}$. If we assume $0 < \mu < 1$, then $QA > 0$ and it follows from the Perron-Frobenius Theorem that from any initial distribution, the process converges to a unique limit $\mathbf{v}^\mu$, the principal right eigenvector of QA. The population determined by $\mathbf{v}^\mu$ marks the culmination of the evolutionary process and is the quasispecies for the evolution described by μ, L and a.

If μ is close to 0 then one would expect $\mathbf{v}^\mu$ to have its mass concentrated on sequences of high fitness but not entirely on the fittest or *master* sequence. On the other hand note that if $\mu = 1/2$, then $v^\mu_\sigma = 1/2^L$ for all σ. This conforms with the intuition that if each bit, during copying, can be 0 or 1 with probability $1/2$, then no genetic information can be retained by the quasispecies. Interestingly, Eigen and his coworkers [ES77, EMS88, EMS89] observed that when the fitness landscape is single-peaked, i.e., the master sequence has fitness above 1 while the rest have fitness 1, this phenomena occurs much below $\mu = 1/2$, in fact around $1/L$. This led them to hypothesize that there exists a particular mutation rate (below $1/2$) beyond which all the 2^L strings become nearly-equally abundant. They called this critical mutation rate the *error threshold* as, beyond this rate the master sequence's genetic information is lost, and called this transition an *error catastrophe*.

Since $\mathbf{v}^\mu$ will hardly ever be $\mathscr{U}$, the uniform distribution on $\{0,1\}^L$, the goal is to find the smallest μ such that $\mathbf{v}^\mu$ is *close* to $\mathscr{U}$. Thus, to define the error threshold one first needs a function that measures closeness. We use the most prevalent way to measure distance between probability distributions: $\|\mathbf{v}^\mu - \mathscr{U}\|_1 := \sum_{\tau\in\{0,1\}^L} |v^\mu_\tau - 1/2^L|$. With this, we define $\mu_\star(\varepsilon) := \min\{\mu \in (0,1) : \|\mathbf{v}^\mu - \mathscr{U}\|_1 \leq \varepsilon\}$.[2] When $\mu = 1/2$, the steady state vector $\mathbf{v}^\mu$ is *exactly* $\mathscr{U}$. Hence, $\mu_\star(\varepsilon) \leq 1/2$. Thus, it is sufficient to focus on the error-rate regime $(0, 1/2]$.

[2] Other notions such as $\|\mathbf{v}^\mu - \mathscr{U}\|_2, \|\mathbf{v}^\mu - \mathscr{U}\|_\infty$ or the difference in Shannon entropies of $\mathbf{v}^\mu$ and $\mathscr{U}$, can also be studied and $\mu_\star$ may change accordingly.

Interlude: The Error Threshold and an Anti-Viral Strategy. The notion of error catastrophe is important in antiviral drug-design as several important viruses are RNA-viruses and their evolution can be captured to the first order by the quasispecies model. From the virus' point of view, a high mutation rate implies greater diversity, which in turn could mean greater adaptability and greater ability to escape the hosts immune responses and pressure from drug therapy. At the same time, too high a mutation rate induces a loss of genetic information. Thus, if we could increase the mutation rate past the error threshold, we would severely compromise the virus' identity. Intriguingly, this strategy is already employed by the body which can produce antibodies that increase mutation rate. Artificially, this effect can also be accomplished by mutagenic drugs such as ribavirin, see [CCA01, MHH$^+$11]. In this setting, knowing the error threshold to a high degree of precision is critical: Inducing the body with excess mutagenic drugs could have undesired ramifications which could lead to complications such as cancer, where as increasing the rate while keeping it below the threshold can increase the fitness of the virus, making it more lethal.

Mathematics of the Quasispecies Model. Since the usefulness of the quasispecies model depends on the existence of the error threshold, the first problem is: Given L, a, and ε, analytically determine the smallest value of μ for which the vector $\mathbf{v}^\mu$ comes ε-close to the uniform distribution in the ℓ_1 norm. The matrix Q is well studied in TCS and is often referred to as the *noisy hypercube* matrix. *All* its eigenvalues and eigenvectors can be written explicitly in terms of the parameters μ and L. Additionally, A is a diagonal matrix whose eigenvectors and eigenvalues are explicit. Hence, it may be tempting to believe that the eigenvalues and eigenvectors for QA can be easily determined. The problem is that A is not explicit (it is given as input) and the fact that it is a diagonal matrix is superflous (we could always make either matrix diagonal). Hence, since the eigenspaces interact in a complex manner, the largest eigenvector of QA does not seem to be derivable in closed form from the spectral data of Q and A. In short, the Boolean function a, which gives the quasispecies model the ability to elegantly capture complex evolutionary interactions, turns out to be the reason this process is difficult to analyze.

While there is abundant empirical proof that the error threshold exists for almost all Boolean/fitness functions of biological relevance, proving this formally has remained an important open problem. Indeed, it is easy to construct examples, even in the single-peaked case, of fitness functions where the error threshold is arbitrarily close to $1/2$. Thus, in order to reinforce the use of the quasispecies model in realistic biological settings where the error threshold is observed much below $1/2$, an important goal is to identify a class of fitness functions that are biologically relevant and rigorously prove the existence of error threshold. One such study has been done recently in [Vis12a]. Here, a class of fitness landscapes which capture a wide variety of practical settings, called (c,k)-finite and bounded, are introduced and the existence of error thresholds established. These fitness landscapes have at most k sequences whose fitness is more than the minimum fitness (1), but remains bounded by c. The error threshold turns out to be about $1/L \cdot \ln(ck)$ and, as one would expect, deteriorates with c and k. Interestingly the proof relies on writing A in the basis corresponding to eigenvectors of Q and using elementary ideas from Fourier analysis and linear algebra.

In the absence of analytic bounds on the error threshold one has to resort to computer simulations. Here, one strategy is to start with a very small value of μ and increase it slowly until the dominant right eigenvector of QA is close to uniform. For this algorithm to be finite, we must make finite increments in μ. But how small must the increment be to ensure we do not miss the error threshold? This strategy implicitly assumes that once the error threshold is reached and the mutation rate is increased beyond it, the eigenvector does not magically reorganize and become far from the uniform distribution, i.e., the distance between the uniform distribution and the dominant right eigenvector of QA is non-increasing. Thus, the next problem is the following: Does $\|\mathbf{v}^\mu - \mathscr{U}\|_1$ go monotonically to 0 as μ goes to $1/2$?

The next question concerns the spectral ratio of the QA matrix which is defined as the ratio of the second-largest to the largest eigenvalue of QA. The spectral ratio can be shown to capture the *convergence rate* or the *speed of evolution* and, hence, is an important parameter from a biological perspective. For instance, when modeling the effect of a mutagenic drug, the convergence rate determines the minimum required duration of treatment. Formally, at time t, the distance $\|\mathbf{m}^t - \mathbf{v}^\mu\|_1$ can be shown to go down geometrically with respect to the spectral ratio. One tempting conjecture is that for $\mu \in (0, 1/2)$, the spectral ratio is at most $1-2\mu$, the spectral ratio of the noisy hypercube. This is false in general but can be shown easily when the fitness function is a tensor. The final problem on quasispecies is: Establish bounds on the spectral ratio.

Asexual Evolution in Finite Populations. Despite several remarkable advances, important gaps remain between the quasispecies model and the realistic evolution of (haploid) asexual populations. Whereas the quasispecies model assumes an infinite population size and, hence, adopts a deterministic approach, real populations are often small enough to lend themselves to substantial stochastic effects. For instance, the effective population size of HIV-1 in infected individuals is about $10^3 - 10^6$ [KAB06, BSSD11], which is believed to underlie the strongly stochastic nature of its evolution. This stochasticity results in the empirical observation of the error threshold varying with population size which has implications in mutagenic drug administration for HIV-1 [TBVD12]. There have been numerous attempts to construct finite population models to serve as refinements of the quasispecies model, [Wil05]. However, most have fallen short since they either do not converge to the quasispecies model as the population goes to infinity or the arguments establishing convergence lack rigor. In a recent study, [DSV12] considered a simple, population-genetics-based model that augments the quasispecies model with randomness and an additional population-size parameter N in the simplest possible way: At each time-step t, N_t^σ denotes be the number of sequences (a random variable) of type σ the total population, $\sum_\sigma N_t^\sigma$, is constrained to be N. In each time step, the ensuing evolution is as follows:

Reproduction: In the reproduction step, each σ produces a_σ copies of itself, giving rise to an intermediate population $I_t := \sum_{\sigma \in \{0,1\}^L} I_t^\sigma$, where $I_t^\sigma := a_\sigma N_t^\sigma$.

Selection: In the selection step, N sequences are chosen at random without replacement from the intermediate population, resulting in the selection of S_t^σ sequences of type σ where $\sum_{\sigma \in \{0,1\}^L} S_t^\sigma = N \le I_t$.

Mutation: In the mutation step, each selected sequence is mutated with probability μ per bit, giving rise to the next generation of N_{t+1}^σ sequences of type σ, such that $\sum_{\sigma \in \{0,1\}^L} N_{t+1}^\sigma = N$.

This (RSM-) model is best viewed as a Markov chain where the state space is the set of functions $f : \{0,1\}^{2^L} \mapsto \{0,1,\ldots,N\}$ such that $\sum_\sigma f(\sigma) = N$. Thus, the number of states of this Markov chain is $\binom{N+2^L-1}{N}$ which is roughly N^{2^L}. It is easy to show that for any $0 < \mu < 1$ the transition matrix of this Markov chain has a unique stationary vector denoted by π. π is indexed by all f satisfying the property above and $\sum_f \pi(f) = 1$. What is the relation between the RSM model and the quasispecies model? It is immediate from the description that $\mathbb{E}\left[{\mathbf{N}_{t+1}}/{N} | \mathbf{N}_t\right] = \frac{QA\mathbf{N}_t}{\|QA\mathbf{N}_t\|_1}$. Thus, the *expected* one step evolution equation in the RSM model is the same as that in the quasispecies model. This also illustrates what makes the RSM model difficult: division of one random variable by another.

Analogous to the vector $\mathbf{v}^\mu$ in the quasispecies model is the vector $\mathbf{V}^\mu$ which captures the expected fraction of each sequence with respect to π for a fixed N : $\mathbf{V}^\mu := \mathbb{E}_\pi\left[\lim_{t\to\infty} {\mathbf{N}_t}/{N}\right]$. Thus, the error threshold, $\mu_\star(\varepsilon, N)$ is now defined to occur when $\|\mathbf{V}^\mu - \mathscr{U}\|_1 \le \varepsilon$ for the first time as μ is increased from 0. It was shown in [DSV12] that $\mu_\star(\varepsilon, N)$ converges $\mu_\star(\varepsilon)$ as $N \to \infty$, however, it is not known how $\mu_\star(\varepsilon, N)$ varies with N. An important question is to understand the dependence of $\mu_\star(\varepsilon, N)$ on N analytically. Experiments suggest that $\mu_\star(\varepsilon, N) \sim (1 - {O(1)}/{\sqrt{N}})\mu_\star(\varepsilon)$ when the fitness landscape and L are fixed.

The applicability of the RSM and related models to the calculation of error threshold relies on the fact that the corresponding Markov chain mixes rapidly, i.e., in time significantly smaller than the number of states in the Markov chain. However, there has been a lack of mixing time bounds. Since there is no easy way to determine if a chain has mixed by looking at samples, in practice, specific parameters are tracked until they seem to *stabilize*. Theoretical justification is required for this as we know that the Markov chains can get *stuck* and the mixing can be slow; there are striking examples of this, see for instance [MV05, SV07, SV11]. This brings us to the important question, an answer to which is needed in order to justify the use of the simulation techniques employed in developing guidelines for drug design: For what Boolean/fitness functions does the RSM Markov chain mix rapidly? This problem turns out not to be so straightforward even for $L = 1$, a case which was recently settled in [Vis12b] by connecting the mixing time to the spectral ratio of the corresponding quasispecies model.

Conclusion. Through this short overview, which admittedly can only be considered the tip of an iceberg, the role that mathematicians and theoretical computer scientists can play should be clear. Since many problems stated here seem to lie at the boundary of or outside our current knowledge, pursuing them can not only provide useful tools to biologists, but can also enrich the tool-set of mathematicans.

References

[BSSD11] Rajesh Balagam, Vasantika Singh, Aparna Raju Sagi, and Narendra M. Dixit. Taking multiple infections of cells and recombination into account leads to small within-host effective-population-size estimates of HIV-1. *PLoS ONE*, 6(1):e14531, 01 2011.

[CCA01] Shane Crotty, Craig E. Cameron, and Raul Andino. RNA virus error catastrophe: Direct molecular test by using ribavirin. *Proceedings of the National Academy of Sciences*, 98(12):6895–6900, 2001.

[CLPV12] Erick Chastain, Adi Livnat, Christos Papadimitriou, and Umesh Vazirani. Multiplicative updates in coordination games and the theory of evolution. *To appear in Innovations in Theoretical Computer Science (ITCS)*, 2013.

[DSV12] N. Dixit, P. Srivastava, and N. K. Vishnoi. A finite population model of molecular evolution: Theory and computation. *In Journal of Computational Biology, 19(10): pp. 1176–1202*, 2012.

[Eig71] M. Eigen. Selforganization of matter and the evolution of biological macromolecules. *Die Naturwissenschaften*, 58:456–523, 1971.

[Eig93] M. Eigen. The origin of genetic information: viruses as models. *Gene*, 135:37–47, 1993.

[EMS88] M. Eigen, J. McCaskill, and P. Schuster. Molecular quasi-species. *J. Phys. Chem.*, 92:6881–6891, 1988.

[EMS89] M. Eigen, J. McCaskill, and P. Schuster. The molecular quasi-species. *Adv. Chem. Phys.*, 75:149–263, 1989.

[ES77] M. Eigen and P. Schuster. The hypercycle, a principle of natural self-organization. part a: Emergence of the hypercycle. *Die Naturwissenschaften*, 64:541–565, 1977.

[HD98] John Holland and Esteban Domingo. Origin and evolution of viruses. *Virus Genes*, 16:13–21, 1998. 10.1023/A:1007989407305.

[KAB06] Roger D. Kouyos, Christian L. Althaus, and Sebastian Bonhoeffer. Stochastic or deterministic: what is the effective population size of HIV-1? *Trends in Microbiology*, 14(12):507 – 511, 2006.

[LEK$^+$99] L A Loeb, J M Essigmann, F Kazazi, J Zhang, K D Rose, and J I Mullins. Lethal mutagenesis of hiv with mutagenic nucleoside analogs. *Proc Natl Acad Sci U S A*, 96(4):1492–1497, February 1999.

[LPDF08] Adi Livnat, Christos Papadimitriou, Jonathan Dushoff, and Marcus W. Feldman. A mixability theory for the role of sex in evolution. *Proceedings of the National Academy of Sciences*, 105(50):19803–19808, 2008.

[MHH$^+$11] James I. Mullins, Laura Heath, James P. Hughes, Jessica Kicha, Sheila Styrchak, Kim G. Wong, Ushnal Rao, Alexis Hansen, Kevin S. Harris, Jean-Pierre Laurent, Deyu Li, Jeffrey H. Simpson, John M. Essigmann, Lawrence A. Loeb, and Jeffrey Parkins. Mutation of HIV-1 genomes in a clinical population treated with the mutagenic nucleoside kp1461. *PLoS ONE*, 6(1):e15135, 01 2011.

[MV05] Elchanan Mossel and Eric Vigoda. Phylogenetic mcmc algorithms are misleading on mixtures of trees. *Science*, 309(5744):2207–9, Sep 2005.

[SV07] Daniel Stefankovic and Eric Vigoda. Pitfalls of heterogeneous processes for phylogenetic reconstruction. *Sys Bio*, 56(1):113–24, Jan 2007.

[SV11] Daniel Stefankovic and Eric Vigoda. Fast convergence of markov chain monte carlo algorithms for phylogenetic reconstruction with homogeneous data on closely related species. *SIAM J. Discrete Math.*, 25(3):1194–1211, 2011.

[TBVD12] K. Tripathi, R. Balagam, N. K. Vishnoi, and N. Dixit. Stochastic simulations suggest that HIV-1 survives close to its error threshold. *In PLoS Computational Biology, 8(9): e1002684*, 2012.

[Vis12a] N. K. Vishnoi. Making evolution rigorous- the error threshold. *To appear in Innovations in Theoretical Computer Science (ITCS)*, 2013.

[Vis12b] N. K. Vishnoi. Manuscript, 2012.

[Wil05] Claus Wilke. Quasispecies theory in the context of population genetics. *BMC Evolutionary Biology*, 5(1):44, 2005.