Research Statement

Steven A. Frank

I study the natural processes that design organisms. My theoretical work solves particular biological puzzles and synthesizes disciplines. Puzzles include genome conflict and sterility, mitochondria and male disease, and somatic mosaicism and cancer. Syntheses cover conflict and cooperation, immunology and pathogen variability, cancer progression and inheritance, and my current project on the evolutionary design of regulatory control. My work sometimes leads to mathematical expressions that unify analysis. Through focus on symmetry, I clarified the fundamental evolutionary principles of natural selection and social interaction. I also found the general forms of commonly observed probability distributions and scaling relations, which help to interpret natural pattern.

Study of organismal design leads to new biological and mathematical problems. The biology and math perspectives provide inseparable insight. However, it is easier to explain the perspectives separately. Two distinct research statements follow, one for the biology and one for the math.

Biological problems

I discuss puzzles, syntheses, and current projects in the following sections.

Puzzles

I have identified several key puzzles and proposed potential solutions. I list four examples. Each example has stimulated new empirical work.

Mitochondria and male disease.[1] Mitochondria typically transmit through the female line. Mutations that cause disease in males and are nearly neutral in females drift in frequency. By contrast, selection removes mutations that are deleterious in females. That sex-biased selective sieve predicts a widespread association between mitochondria and male disease. Our work identified this puzzle, which had not previously been noted.

This topic has become a mini-discipline, sometimes known as *Mother's curse.* Several empirical studies designed to test this idea found support.[2,3] Numerous theoretical articles have developed the mathematical biology of this topic.

Meiotic drive, Haldane's rule, and speciation.[4] The first step in hybrid species incompatibility is typically sterility of the heterogametic sex, a pattern known as Haldane's rule. I suggested that genomic conflict arising from sex-chromosome meiotic drive may explain this sex-biased pattern of hybrid sterility. Conflicts often associate with rapid evolutionary change. Meiotic drive within species would likely lead to rapid divergence between species (see also Hurst & Pomiankowski [5]).

This idea about genomic conflict and speciation founded a mini-discipline. Several labs now devote their research to this topic. Although the idea was originally controversial, it is now widely accepted as one of the best supported theories for the genetics of speciation.

Infective dose.[6] The number of pathogens required to start an infection varies widely between species. No general theory explained that wide variation. We suggested that the particular molecular virulence mechanisms during initial pathogen invasion may explain variation in infective dose.

We predicted that virulence factors directly injected into neighboring host cells require few initial pathogens to start an infection. By contrast, virulence factors that act distantly on host immune regulation may require many initial pathogens to generate a sufficient concentration of diffusible factors. Local versus global virulence factor action may correspond to small versus large minimum infective dose.

Our work is the only theory to explain the widely varying infective dose observed among pathogens.

Existing data support our theory,[7] but more direct experimental studies are needed. Mathematical models of pathogen invasion dynamics and immune response will help to develop the topic.

Somatic mosaicism.[8] A human body has about 100 trillion cells derived from the single zygote. The vast number of cell divisions introduces many mutations, causing widespread somatic mosaicism. Our theory predicts great diversity in mosaicism between individuals. That mosaic diversity may explain a significant fraction of the variance in predisposition to disease.

We were the first to relate mosaicism to the mathematical theory of branching cellular lineages, with emphasis on cancer risk.[8–10] Our theory provided the first clear predictions about the high level of mosaicism expected within individuals and the great variability in mosaicism and the risk of disease between individuals. My later article provided the first direct connection between neurodegeneration and a fully realized theory of mosaicism.[11]

With the advances in single-cell genomics, this topic has developed into a major research field. So far, empirical studies have mostly confirmed the existence of mosaicism. Going forward, I outlined several key problems that have yet to be analyzed empirically.[12]

Syntheses

My books and series synthesize key topics and set the direction for future work.

Foundations of social evolution.[13] This book unified the mathematical theory of natural selection applied to economic problems of organismal resource allocation, game theory aspects of social cooperation, and population aspects of demography. The mathematical methods that I developed in this book became the standard for much of the subsequent work in the field, leading to many testable empirical predictions for sex allocation, life history, and social behavior. This book has also influenced the mathematical development of evolutionary models in economics.[14]

Immunology and evolution of infectious disease.[15] This book focuses on pathogen variation to escape host immune recognition. It integrates molecular biology, immunology, pathogen biology, and population dynamics within a quantitative framework.

With regard to my general interest in organismal design, the immune system provides an excellent case study. The immune system deploys a variety of search, recognition, and defense tactics, raising interesting problems about the evolutionary design and integration of those components.

Each chapter of the book finishes with a listing of key unsolved research topics. Those topics provide the basis for new research and also present a series of interesting challenges that can be used in teaching. I have also published additional research articles on pathogen variability and immune escape.[16–19]

Dynamics of cancer.[10] This book provides the only comprehensive synthesis of age-specific cancer incidence patterns with mathematical theory for the causes of cancer progression dynamics. To achieve that synthesis, I developed novel quantitative approaches to analyze epidemiological data. For example, my focus on the acceleration of cancer with age naturally develops the duality between acceleration and force, allowing direct study of biological causes.

To develop testable theories about cause, I built a comprehensive mathematical framework for the dynamics of cancer progression. To set the basis for future work, I integrated the epidemiology and mathematical theory of causation with the molecular details of regulatory controls and how those controls break down in cancer.

With regard to my general interest in organismal design, cancer is one of the great subjects of modern biology. As I say in the first paragraphs of the book:

> Through failure we understand biological design. Geneticists discover the role of a gene by studying how a mutation causes a system to fail. Neuroscientists discover mental modules for face recognition or language by observing how particular brain lesions cause cognitive failure.
>
> Cancer is the failure of controls over cellular birth and death. Through cancer, we discover the design of cellular controls that protect against tumors and the architecture of tissue restraints that slow the progress of

disease.

Many opportunities remain to develop new evolutionary and quantitative analyses of cancer and the related topic of regulatory control.[20]

Natural selection. I wrote a series of seven articles on the theory of natural selection.[21–27] In my *Mathematical problems* research statement below, I mentioned related work on fundamental aspects of symmetry, dynamics, and information.

The first article presents my framework for mathematical models of evolution when natural selection varies over time or space.[22,28] That framework remains the standard approach in the current literature for analyzing variable selection in a unified way. My work also develops the connection between the mathematics of natural selection and the mathematics of economic returns under risk and uncertainty.

The second article develops a novel synthesis of how phenotypic variability influences the rate and direction of evolutionary change by natural selection.[21] I used that new theory as the basis for understanding how resistance to cancer therapy evolves.[29] The evolution of drug resistance relates to the more general problem of how new traits arise and spread to meet the challenge of novel and extreme environments. The topic of resistance evolution is very active and provides many opportunities for theory on biomedically relevant problems.

Other articles include a novel analysis of the levels at which natural selection acts,[23] classic causal modeling and path analysis descriptions of natural selection,[26] and a modern summary of kin selection theory along with a historical perspective on that topic.[27] Together, the articles present a comprehensive synthesis of natural selection with regard to problems of organismal design.

Control theory tutorial.[30] I recently taught myself engineering control theory. I am using that theory to develop new projects on the evolutionary design of regulatory control (see below). To teach myself, I wrote a tutorial and software package, which I published as a small book.[30]

The theory of engineering control builds large models of dynamical systems by transforming linear components in the time domain into the complex Laplace domain. The advantage of the transformed expressions is that one can multiply complex Laplace signals expressed by transfer functions through a sequence of processes. That approach allows study of dynamics in large systems with respect to signal frequency and intensity. One can then easily analyze the evolutionary consequences of alternative sensor, filter, and control designs for various tradeoffs in performance measures.

Current projects

The paradox of robustness. The better a system is at correcting errors, the more that system can tolerate mistakes made by its components. Because the mistakes by components of a robust error-correcting system do not matter so much, such systems tend to accumulate variable components that decay in performance. Better error correction begets more errors, the paradox of robustness.[31–33]

The tendency for robustness to cause the accumulation of mutations has been discussed previously.[34–36] However, the paradox of robustness applies much more broadly, because greater robustness at the system level reduces the pressure of natural selection on the performance of the system's components. That change in the intensity of natural selection alters the costs and benefits of component design.

I introduced this broader notion of the paradox of robustness in my prior publications.[31–33] However, the theory has not been developed. No compelling applications to specific systems have been completed.

Beyond my prior work, the leading geneticist Michael Lynch made the strongest case. He emphasized that cells have many layered mechanisms of error correction. He recognized the difficulty of understanding how such layered, protective robustness could evolve based on current theories of evolutionary dynamics.

Lynch[37] said: "If, however, drift prevents natural selection from inexorably moving cellular features toward a state of molecular perfection, how do we account for the abundant examples of organisms using layered mechanisms for dealing with intracellular problems?...As pointed out by Frank,[32] an appreciation for the internal evolutionary dynamics of redundant systems provides an alternative perspective on

the origin and maintenance of the myriad of molecular attributes often interpreted as acquired enhancements of cellular robustness."

Layered error-correcting controls over traits occur throughout cell biology, physiology, behavior, and social systems. No theory explains how such layered controls arise by evolutionary dynamics. My project advances this topic.

As a first step, I am currently developing a control theory analysis for the evolutionary design of regulatory control.

Evolutionary design of regulatory control. Much of modern biology focuses on the molecular mechanisms that regulate biochemical and physiological processes. Progress on mechanism and immediate function raises the problem of how such complex controls evolve by natural processes.

To study the evolutionary design of regulatory control, it is useful to divide the problem into two steps: the general abstract theory and the application of that theory to particular biological systems.

The work must begin with the basic abstract theory. If we do not understand, even in simple theories, how evolutionary processes may shape system design, then we certainly cannot understand the details of particular systems. As the basic theoretical work progresses, the challenge will shift to application. How can we use the abstract theory to make testable predictions about real systems?

Currently, I am focusing on the first step of abstract theory. How does natural selection sort among the many tradeoffs in performance that shape the broad features of regulatory control? How does the evolutionary design of regulatory control architecture influence broadly observable patterns, such as genetic variability of system components and stochasticity of trait expression between individuals?

An abstract evolutionary theory of regulatory control is not a trivial problem. Consider error-correcting feedback, perhaps the single greatest principle of control system design in both human-engineered and biological systems.

In an error-correcting feedback system, the error measures the difference between a system's actual output and its target. By feeding back the error as an input, the system can move in the direction that reduces the error. Error correction compensates robustly for misinformation about system dynamics and for perturbations to system components. Excellent performance often follows in spite of limited information, sloppy components, and noisy signals.

A robust error-correcting feedback system compensates for sloppy, error-prone components. That robust compensation weakens the pressure of natural selection on the components, the paradox of robustness. Thus, the evolution of each additional error-correcting feedback loop at the system level will tend to associate with the evolution of cheaper, lower performing system components. Those components may also tend to accumulate greater genetic variability and stochasticity of expression.

To build a theoretical framework, I am working toward a series of articles.

- Design tradeoffs and control theory: combining evolutionary analysis with engineering control theory provides the essential methods.[38]
- Genetic variability and stochasticity of trait expression: the paradox of robustness increases component variability and the heritability of disease.[39]
- Decay of costly components: the paradox of robustness favors substitution of cheaper, lower performing components within systems.
- Learning as a robustness mechanism: systems that acquire information and adjust control have an additional robustness layer with further consequences from the paradox of robustness.
- Wiring of control architecture: the evolutionary process of building layered control architectures yields seemingly haphazard, complex wiring of control.

After developing the initial theory, the challenge then becomes how to turn the abstract theory into testable predictions for specific systems. One likely path is to find simple regulatory control systems that vary in architecture between closely related populations or species. Microbial systems often provide the best opportunities.

Additionally, the theory may potentially be applied to the extensive modern datasets on genetic variability and single-cell stochasticity of gene expression. In

principle, it should be possible to make comparative predictions about the relative levels of variability of particular genes in relation to the function of those genes within particular regulatory control architectures.

Common patterns: invariance and scale. Below, I discuss my mathematical theory of common probability patterns. To extend that topic, I am looking for applications to biology. How can the abstract theory of common patterns help to understand natural phenomena? To give a sense of possible directions for this work, I briefly summarize my recent article on perception.[40]

The probability that an organism perceives two stimuli as similar typically decays exponentially with separation between the stimuli. The exponential decay in perceptual similarity is often referred to as the universal law of generalization.[41,42]

Both theory and empirical analysis depend on the definition of the perceptual scale. For example, how does one translate the perceived differences between two circles with different properties into a quantitative measurement scale?

There are many different suggestions in the literature for how to define a perceptual scale. Each of those suggestions develops very specific notions of measurement based, for example, on information theory, Kolmogorov complexity theory, or multidimensional scaling descriptions derived from observations.[41–43]

I showed that the inevitable shift invariance of any reasonable perceptual scale determines the exponential form for the universal law of generalization in perception.[40] All of the other details of information, complexity, and empirical scaling are superfluous with respect to understanding why the universal law of generalization has the exponential form.

In some cases, the probability of perceived similarity is Gaussian rather than exponential. I showed that, when the separation between stimuli depends on various underlying perceptional dimensions, it sometimes makes sense to assume that the perceptual scale will also obey exchangeability or rotational invariance. When that additional invariance holds, the universal law takes on the Gaussian form.[40,44]

The exponential and Gaussian forms are particular expressions of the canonical form for probability patterns that I presented in eqn 6. However, not all commonly observed patterns are exponential or Gaussian. Other patterns arise through scaling relations between observed measurements and meaningful shift-invariant scales, represented by w in eqn 6.

The interesting problem, for both biology and mathematics, is how to understand the genesis of various forms for w in different applications. I gave a couple of examples in recent articles.[45,46] But the more general problem remains unsolved. Perhaps there is some way to understand the commonly observed macroscopic scaling relations as asymptotic functional forms when aggregated over the microscopic variability in functional relations.

Mathematical problems

My studies of organismal design and evolutionary process have often led to conceptual challenges. How can we understand the fundamental principles of natural selection? When studying the mathematical theory of natural selection, how can we understand the deeper relations between that theory and other apparently similar mathematical theories from different disciplines? How does deeper abstract mathematical understanding improve our ability to analyze evolutionary problems?

I have also faced the difficulty of understanding why certain common patterns recur in data and in predictions from theory. Those common patterns typically arise as probability distributions or scaling relations. To evaluate the causes of natural pattern, I had to struggle with the general theory of commonly observed patterns. That struggle led me to new mathematical work on the common patterns of nature.

Natural selection

Natural selection may be described abstractly as the change in some characteristics of a population in response to a force. We can think of a population as a probability distribution of characteristics. The force of natural selection drives the evolutionary dynamics of change in probability distributions.

Many of the particular properties of natural section and the evolutionary dynamics of probability

distributions arise from simple underlying invariances, or symmetries. For example, the conservation of total probability, the summing of all probabilities to one, imposes a universal invariance on changes in populations and on the dynamics of probability distributions.

Once one recognizes the universal invariance underlying the evolutionary change in populations, one can show the unity of natural selection, information theory, entropy, the forms of common probability distributions, and classic descriptions of dynamics in physics.[47] Here, I summarize a few results from my past work.

Set mapping for evolutionary change

Start with a population as a set of things. Each thing has a property indexed by i. Those things with a common index comprise a fraction, q_i, of the population and have average value, z_i, for whatever we choose to measure by z. Write $\mathbf{q}$ and $\mathbf{z}$ as the vectors over all i. The population average value is $\bar{z} = \mathbf{q} \cdot \mathbf{z}$.

A second population has matching vectors $\mathbf{q}'$ and $\mathbf{z}'$. For frequency, $q_i' = w_i q_i$, in which w_i describes frequency change and, in biology, is realized relative fitness. Here, q_i' is the fraction of the second population derived from entities with index i in the first population, a set mapping. Likewise, z_i' is the average value in the second population of members derived from entities with index i in the first population. Let Δ be the difference between the derived population and the original population.

We can write the abstract description for the change in average value as[25]

$$\Delta\bar{z} = \Delta(\mathbf{q} \cdot \mathbf{z}) = \Delta\mathbf{q} \cdot \mathbf{z} + \mathbf{q}' \cdot \Delta\mathbf{z}. \tag{1}$$

To express this description in terms of the forces acting on frequency change, we use the above $q_i' = w_i q_i$ to define

$$a_i = w_i - 1 = \frac{q_i'}{q_i} - 1 = \frac{\Delta q_i}{q_i}, \tag{2}$$

which, in biology, is Fisher's average excess in fitness.

We can use any values for $\mathbf{z}$. Choose $\mathbf{z} \equiv \mathbf{a}$. Then

$$\Delta\bar{a} = \Delta(\mathbf{q} \cdot \mathbf{a}) = \Delta\mathbf{q} \cdot \mathbf{a} + \mathbf{q}' \cdot \Delta\mathbf{a} = 0. \tag{3}$$

The equality to zero expresses the conservation of total probability

$$\bar{a} = \mathbf{q} \cdot \mathbf{a} = \sum_i q_i \frac{\Delta q_i}{q_i} = \sum_i \Delta q_i = 0,$$

because the total changes in probability must cancel to keep the sum of the probabilities constant at one. Thus, eqn 3 appears as a seemingly trivial result, a notational spin on $\sum \Delta q_i = 0$. However, many generalities of the genetical theory of natural selection follow from the partition of conserved probability into the two terms of eqn 3. In addition, the partition unifies many formal relations between natural selection and information theory, the dynamics of entropy and probability, and basic aspects of physical dynamics.[47]

Force and inertial frame in mechanics

The power of eqn 3 derives from its partition of the balancing components of change into two parts, $\Delta\mathbf{q} \cdot \mathbf{a}$ and $\mathbf{q}' \cdot \Delta\mathbf{a}$. With a bit more notational manipulation, we arrive at an abstract nondimensional analogy of D'Alembert's principle of mechanics for conservative systems[47]

$$\Delta\bar{a} = (\mathbf{F} + \mathbf{I}) \cdot \Delta\mathbf{q} = 0, \tag{4}$$

which describes the balance between the change by direct forces, $\mathbf{F}$, and the change with respect to acceleration in the inertial frame of reference, $\mathbf{I}$.

D'Alembert generalizes Newton's force equals mass times acceleration to multiple dimensions and, here, to a more abstract interpretation. According to Lanczos,[48] the power of D'Alembert's partition is that it "focuses attention on the forces, not on the moving body..." In the analysis of complex dynamics, it often helps to focus on abstract notions of force that drive system change, such as fitness or entropy production. I mention three examples.

Fisher's fundamental theorem

The most famous result in population genetics theory, Fisher's fundamental theorem of natural selection, follows immediately. The first terms of eqns 3

and 4 yield

$$\begin{aligned}\Delta\bar{a} = \Delta\mathbf{q}\cdot\mathbf{a} = \Delta\mathbf{q}\cdot\mathbf{F} &= \sum_i \Delta q_i\left(\frac{\Delta q_i}{q_i}\right)\\ &= \sum_i q_i\left(\frac{\Delta q_i}{q_i}\right)^2 = \sum_i q_i a_i^2\\ &= V_w,\end{aligned}$$

which shows that the rate of change in average fitness, $\Delta\bar{a}$, caused by the direct force of natural selection, equals the variance in fitness, V_w.

Here, I described fitness by the average excess, a. Fisher's theorem actually states that the rate of change in fitness caused by the direct force of natural selection is equal to the variance in the average effects of fitness. We easily obtain the average effects as the partial regressions of fitness on some predictors, such as various genes. We can then substitute the average effects directly into the equations above to get Fisher's theorem[49] (not shown here). The important conceptual point is that we have a generalization of Fisher's partition of total change in fitness into direct and inertial components, focusing on the forces and not on the "moving bodies," or moving gene frequencies.

We can generalize Fisher's theorem to the change in any value or quantitative character that we assign to entities. Simply make a change of coordinates $\mathbf{a} \mapsto \mathbf{z}$, getting us back to eqn 1 in a way that we can use what we learned from studying the invariant, conserved form of eqn 3. That generalization allowed me to unify the theory of social evolution.[13,27]

I discussed interpretations of the inertial component of total change and the analysis of nonconservative systems in Frank [47].

Dynamics of information

We can connect classic mathematical models of natural selection and evolutionary dynamics to classic expressions for changes in information. The key arises from my abstract partition of change between sets, constrained by invariant total probability.

Start with the direct force component of total change

$$\Delta\mathbf{q}\cdot\mathbf{a} = \sum_i \Delta q_i\left(\frac{\Delta q_i}{q_i}\right).$$

For small Δq_i, we can write

$$a_i = \frac{\Delta q_i}{q_i} \to \log\frac{q_i'}{q_i}, \tag{5}$$

thus[24]

$$\Delta\mathbf{q}\cdot\mathbf{a} = \sum_i (q_i' - q_i)\log\frac{q_i'}{q_i} = \mathcal{D}\left(q'||q\right) + \mathcal{D}\left(q||q'\right),$$

in which $\mathcal{D}$ is the Kullback-Leibler divergence, the most fundamental measure for the change in information from classic information theory.[50] For small Δq_i, the value of $\mathcal{D}$ is the Fisher information metric, the foundation for information geometry[51] and much of the classic theory of statistical inference.[52]

The variance in fitness, V_w, from Fisher's fundamental theorem of natural selection, is better understood as the divergence or distance between two sets. In biology, the sets are ancestral and descendant populations. The separation between populations, or sets, can be described by classic measures of information theory.

We may say that the direct force of natural selection causes populations to accumulate information about the environment equal to the sum of the forward and backward $\mathcal{D}$ measures.[24,47,53] That sum is also known as the Jeffreys divergence. In the limit of small changes, the Jeffreys divergence becomes the Fisher information metric.

The match between selection and information follows from the simple underlying invariance of conserved probability and the partition of that invariant quantity into two terms, matching D'Alembert's partition. This abstract generalization clarifies the wide variety of vague statements about how natural selection, information, statistical inference, and classic mechanics relate to each other.

Maximum entropy probability distributions

Many problems in biology turn on understanding the genesis of particular probability distributions. In my work on cancer, I noted the strong tendency for age of cancer onset patterns and mortality patterns to match a gamma distribution or one of the extreme value distributions.[10,45] Variants of power-law patterns often arise in biological data.[46] How can we understand those common patterns of nature?[54]

Jaynes[55] argued that we can understand commonly observed probability distributions by supposing that dynamics tends to maximize entropy subject to constraints. Jaynes sought to overthrow Boltzmann's canonical ensemble for statistical mechanics. The canonical ensemble describes macroscopic probability patterns by aggregation over a large number of equivalent microscopic particles.

The theory of statistical mechanics, based on the microcanonical ensemble, yields several commonly observed probability distributions. However, Jaynes emphasized that the same probability distributions commonly arise in economics, biology, and many other disciplines. In those nonphysical disciplines, there is no meaningful canonical ensemble of identical microscopic particles. According to Jaynes, there must another more general cause of the common probability patterns. The maximization of entropy is one possibility.

In Frank [47], I showed that the fundamental expression of change in eqn 1 includes Jaynesian maximum entropy as a special case.

From eqn 2, we can write $\Delta q_i = q_i a_i$. The condition for equilibrium with regard to frequencies is $a_i = 0$, or, from eqn 5, $\log q_i = \log q_i'$. Noting that $q_i' = q_i + \Delta q_i$, a constraint on the vector of frequency changes, $\Delta\mathbf{q}$, will constrain the equilibrium probability distribution, $\mathbf{q}^*$. In eqn 1, suppose that $\Delta\bar{z} = 0$, which means that the average value, $\bar{z} = \mathbf{q}\cdot\mathbf{z}$, is constant. That invariance of average value constrains the pattern of change in frequencies.

It turns out that the constraint on average value can be expressed by[47]

$$\log q_i' = \log \tilde{k}_i - \lambda z_i,$$

in which the $\tilde{k}_i$ are constants chosen to satisfy the conservation of total probability. At equilibrium, the probability distribution, $\mathbf{q}^*$, is

$$q_i^* = ke^{-\lambda z_i}.$$

We can match this equilibrium probability distribution to a Lagrangian for the dynamics associated with eqn 1 as

$$\mathcal{L} = \mathcal{E} + \tilde{k}\left(\sum q_i - 1\right) - \lambda\left(\sum q_i z_i - \mu\right),$$

in which the first term, $\mathcal{E} = -\sum q_i \log q_i$, is the classic definition of information entropy, the second term is the constraint on total probability, and the third term is the constraint on the mean value, $\bar{z}$. This Lagrangian is Jaynes' expression for how to obtain maximum entropy probability distributions subject to constraint. Alternative constraints yield alternative probability distributions. I showed that Jaynes' results follow from the basic abstract expression for change in populations.[47]

I used the Jaynesian maximum entropy framework to study common probability patterns in biological problems.[54,56] However, it was not clear how to develop meaningful constraints to match the variety of commonly observed probability distributions. Also, Jaynesian maximum entropy does not reveal the deeper relations between the different forms of commonly observed distributions. To understand those problems, I looked for the simple underlying symmetries that unify understanding of common probability patterns.

Common probability patterns

We can obtain the canonical form for nearly all common continuous probability distributions from a few simple invariances.[44] Suppose the probability distribution function (pdf) for a continuous variable y is $f(T_y)$, in which T_y is a function of y that defines the "natural" measurement scale. For example, y may have a natural logarithmic scaling, $T_y = \log(1 + y)$.

Assume that a natural scale, T_y, means that the associated pdf is affine invariant to a constant shift and a constant stretch, $T_y \mapsto a + bT_y$, that is, $f(a + bT_y) \mapsto f(T_y)$. From these assumptions, I showed[44] that the pdf value associated with y is

$$q_y = f(T_y) = ke^{-\lambda T_y},$$

in which the constant k is set by the conservation of total probability, associated with shift invariance, and the constant λ is set by the conservation of average value, associated with stretch invariance.[44]

Affine invariance is often a sensible requirement for a natural scale. For example, suppose that, associated with y, we have measurements of temperature on the Celsius scale, T_y. Then we would expect that transforming to the Fahrenheit scale, $\tilde{T}_y = 32 + 1.8T_y$, would leave the associated probability pattern unchanged.

Affine invariance of the pdf with respect to T_y implies additional structure. Write the affine invariance

in terms of a generator process, $G(y)$, such that each application of the generator leaves the pdf invariant, because

$$T[G(y)] = T \circ G = a + bT_y.$$

After n applications of G, the probability pattern remains unchanged. An infinitesimal application of the generator also retains the invariant pdf form. We can write the infinitesimal transformation as a differential equation with respect to a base scale,[57] $w(y)$, as

$$T \circ G^{\epsilon} = \frac{\mathrm{d}T}{\mathrm{d}w} = \alpha + \beta T,$$

which, dropping shift and stretch constants, has the solution

$$T = e^{\beta w},$$

with $T \to w$ as $\beta \to 0$ when accounting for the shift and stretch factors that are not shown here. Using this general form for T, we obtain the canonical expression that describes nearly all commonly observed continuous probability distributions[44,45]

$$q\,\mathrm{d}\psi = ke^{-\lambda e^{\beta w}}\mathrm{d}\psi, \tag{6}$$

when we add a few additional details about the measure, $\mathrm{d}\psi_y$, and the commonly observed base scales, $w(y)$. Understanding the abstract form of common probability patterns clarifies the study of many biological problems.[40,45,46]

Summary

The results in this mathematical section can be thought of as a broad framing of conjectures. The conjectures show the potential for a symmetry-based unification of diverse scientific and mathematical topics.

References

[1] Frank, SA & Hurst, LD 1996. Mitochondria and male disease. *Nature* 383, 224.

[2] Frank, SA 2012. Mitochondrial burden on male health. *Current Biology* 22, R797–R799.

[3] Patel, MR et al. 2016. A mitochondrial DNA hypomorph of cytochrome oxidase specifically impairs male fertility in *Drosophila melanogaster*. *eLife* 5, e16923.

[4] Frank, SA 1991. Divergence of meiotic drive-suppression systems as an explanation for sex-biased hybrid sterility and inviability. *Evolution* 45, 262–267.

[5] Hurst, LD & Pomiankowski, A 1991. Causes of sex ratio bias may account for unisexual sterility in hybrids: a new explanation of Haldane's rule and related phenomena. *Genetics* 128, 841–858.

[6] Schmid-Hempel, P & Frank, SA 2007. Pathogenesis, virulence, and infective dose. *PLoS pathogens* 3, e147.

[7] Leggett, HC, Cornwallis, CK & West, SA 2012. Mechanisms of pathogenesis, infective dose and virulence in human parasites. *PLoS Pathogens* 8, e1002512.

[8] Frank, SA & Nowak, MA 2003. Developmental predisposition to cancer. *Nature* 422, 494.

[9] Frank, SA 2003. Somatic mosaicism and cancer: inference based on a conditional Luria-Delbruck distribution. *Journal of Theoretical Biology* 223, 405–412.

[10] Frank, SA 2007. *Dynamics of Cancer: Incidence, Inheritance, and Evolution*. Princeton, NJ: Princeton University Press.

[11] Frank, SA 2010. Somatic evolutionary genomics: mutations during development cause highly variable genetic mosaicism with risk of cancer and neurodegeneration. *Proceedings of the National Academy of Sciences USA* 107, 1725–1730.

[12] Frank, SA 2014. Somatic mosaicism and disease. *Current Biology* 24, R577–R581.

[13] Frank, SA 1998. *Foundations of Social Evolution*. Princeton, New Jersey: Princeton University Press.

[14] Andersen, ES 2004. Population thinking, Price's equation and the analysis of economic evolution. *Evolutionary and Institutional Economics Review* 1, 127–148.

[15] Frank, SA 2002. *Immunology and Evolution of Infectious Disease*. Princeton, NJ: Princeton University Press.

[16] Barbour, AG, Dai, Q, Restrepo, BI, Stoenner, HG & Frank, SA 2006. Pathogen escape from host immunity by a genome program for antigenic variation. *Proceedings of the National Academy of Sciences USA* 103, 18290–18295.

[17] Frank, SA 1999. A model for the sequential dominance of antigenic variants in African trypanosome infections. *Proceedings of the Royal Society of London B* 266, 1397–1401.

[18] Frank, SA & Barbour, AG 2006. Within-host dynamics of antigenic variation. *Infection, Genetics and Evolution* 6, 141–146.

[19] Frank, SA & Bush, RM 2007. Barriers to antigenic escape by pathogens: trade-off between reproductive rate and antigenic mutability. *BMC Evolutionary Biology* 7, 229.

[20] Frank, SA 2016. Commentary: The nature of cancer research. *International Journal of Epidemiology* 45, 638–645.

[21] Frank, SA 2011. Natural selection. II. Developmental variability and evolutionary rate. *Journal of Evolutionary Biology* 24, 2310–2320.

[22] Frank, SA 2011. Natural selection. I. Variable environments and uncertain returns on investment. *Journal of Evolutionary Biology* 24, 2299–2309.

[23] Frank, SA 2012. Natural selection. III. Selection versus transmission and the levels of selection. *Journal of Evolutionary Biology* 25, 227–243.

[24] Frank, SA 2012. Natural selection. V. How to read the fundamental equations of evolutionary change in terms of information theory. *Journal of Evolutionary Biology* 25, 2377–2396.

[25] Frank, SA 2012. Natural selection. IV. The Price equation. *Journal of Evolutionary Biology* 25, 1002–1019.

[26] Frank, SA 2013. Natural selection. VI. Partitioning the information in fitness and characters by path analysis. *Journal of Evolutionary Biology* 26, 457–471.

[27] Frank, SA 2013. Natural selection. VII. History and interpretation of kin selection theory. *Journal of Evolutionary Biology* 26, 1151–1184.

[28] Frank, SA & Slatkin, M 1990. Evolution in a variable environment. *American Naturalist* 136, 244–260.

[29] Frank, SA & Rosner, MR 2012. Nonheritable cellular variability accelerates the evolutionary processes of cancer. *PLoS Biology* 10, e1001296.

[30] Frank, SA 2018. *Control Theory Tutorial: Basic Concepts Illustrated by Software Examples*. Cham, Switzerland: Springer.

[31] Frank, SA 2004. Genetic variation in cancer predisposition: mutational decay of a robust genetic control network. *Proceedings of the National Academy of Sciences USA* 101, 8061–8065.

[32] Frank, SA 2007. Maladaptation and the paradox of robustness in evolution. *PLoS ONE* 2, e1021.

[33] Frank, SA 2013. Evolution of robustness and cellular stochasticity of gene expression. *PLoS Biology* 11, e1001578.

[34] Rutherford, SL & Lindquist, S 1998. Hsp90 as a capacitor for morphological evolution. *Nature* 396, 336–342.

[35] Visser, J de, Hermisson, J, Wagner, GP, Meyers, LA, Bagheri-Chaichian, H, Blanchard, JL, Chao, L, Cheverud, JM, Elena, SF, Fontana, W, Gibson, G, Hansen, TF, Krakauer, D, Lewontin, RC, Ofria, C, Rice, SH, Dassow, G von, Wagner, A & Whitlock, MC 2003. Perspective: evolution and detection of genetic robustness. *Evolution* 57, 1959–1972.

[36] Wagner, A 2013. *Robustness and Evolvability in Living Systems*. Princeton University Press.

[37] Lynch, M 2012. Evolutionary layering and the limits to cellular perfection. *Proceedings of National Academy Sciences USA* 109, 18851–18856.

[38] Frank, SA 2018. Evolutionary design of regulatory control. I. A robust control theory analysis of tradeoffs. *bioRxiv*. DOI: 10.1101/332999.

[39] Frank, SA 2018. Evolutionary design of regulatory control. II. Robust error-correcting feedback increases genetic and phenotypic variability. *bioRxiv*. DOI: 10.1101/405456.

[40] Frank, SA 2018. Measurement invariance explains the universal law of generalization for psychological perception. *Proceedings of National Academy of Sciences USA* 115, 9803–9806. DOI: 10.1073/pnas.1809787115.

[41] Chater, N & Vitányi, PM 2003. The generalized universal law of generalization. *Journal of Mathematical Psychology* 47, 346–369.

[42] Shepard, RN 1987. Toward a universal law of generalization for psychological science. *Science* 237, 1317–1323.

[43] Sims, CR 2018. Efficient coding explains the universal law of generalization in human perception. *Science* 360, 652–656.
[44] Frank, SA 2016. Common probability patterns arise from simple invariances. *Entropy* 18, 192.
[45] Frank, SA 2016. Invariant death. *F1000Research* 5, 2076.
[46] Frank, SA 2016. The invariances of power law size distributions. *F1000Research* 5, 2074.
[47] Frank, SA 2017. Universal expressions of population change by the Price equation: Natural selection, information, and maximum entropy production. *Ecology and Evolution* 7, 3381–3396.
[48] Lanczos, C 1986. *The Variational Principles of Mechanics*. 4th ed. New York: Dover Publications.
[49] Frank, SA 1997. The Price equation, Fisher's fundamental theorem, kin selection, and causal analysis. *Evolution* 51, 1712–1729.
[50] Cover, TM & Thomas, JA 1991. *Elements of Information Theory*. New York: Wiley.
[51] Amari, S & Nagaoka, H 2000. *Methods of Information Geometry*. New York: Oxford University Press.
[52] Fisher, RA 1925. Theory of statistical estimation. *Math. Proc. Cambridge Phil. Soc.* 22, 700–725.
[53] Frank, SA 2009. Natural selection maximizes Fisher information. *Journal of Evolutionary Biology* 22, 231–244.
[54] Frank, SA 2009. The common patterns of nature. *Journal of Evolutionary Biology* 22, 1563–1585.
[55] Jaynes, ET 2003. *Probability Theory: The Logic of Science*. New York: Cambridge University Press.
[56] Frank, SA 2014. How to read probability distributions as statements about process. *Entropy* 16, 6059–6098.
[57] Frank, SA & Smith, E 2011. A simple derivation and classification of common probability distributions based on information symmetry and measurement scale. *Journal of Evolutionary Biology* 24, 469–484.