Mendelian Dynamics and Sturtevant's Paradigm .

M. Gromov.

June 26, 2007

Abstract

This a brief introduction to the formal genetics centered around two mathematical ideas going back to Gregor Mendel and Alfred Sturtevant.

Preambule. The road from biology (or from any branch of science on the fundamental level) to mathematics goes in several (often Brownian rather than straight) paths in parallel.

• Identifying a class of phenomena–"particular trees in a forest"–that appear with a regularity suggesting an underlying (mathematical) structure. (Are there non-mathematical structures?)

•• Designing and performing experiments/observations purifying and amplifying what is seen by the naked eye (e.g. by planting our "trees" to an "artificial soil").

••• Making (often implicitly) ad hoc hypotheses, (e.g. continuity, symmetry, functoriality)– that provide a logical framework for the experimental data.

For example, the "theory of coin tossing" derives its mathematical beauty and the (probabilistic) predictive power not from such "definitions" as "the probability is a measure of uncertainty" but from the *assumption* that the the probability distribution on the space $\mathbb{Z}_2^n$ of the *imaginary* outcomes (binary n-sequences) equals the (normalized) *Haar* measure that is, moreover, *invariant under the permutation group* S_n.

Such hypotheses (assumptions), fragments of the grammer of the language in which Nature delievers her messages, are what a mathematician is primerly interested in, while a scientist is concerned with the "meaning" of a message–the structure that is harder to formilize.

Desiphering the grammer of Nature, or, biologically speaking, guessing the design of a seed by looking at (sample branches of) the grown tree, is rarely (if ever) done by mathematicians (Newtons do not count), even when the experimental data are abundant (as in the present day molecular biology). The past mathematical experience channals your imagination toward the old rather than new mathematical concepts.

Even rigorously reformulating such hypotheses is not a straighforward task. For example, the Dirac δ-function needed the theory of distributions to be accepted by mathematicians and the functoriality of the (derivation of the) Boltzmann equation was recognized (albeit not much exploited till now) only with the advent of the "functoriality paradigm" within pure mathematics.

When a "seed" is cultivated in a "mathematical soil", what grows out of it– mathematician's tree–might look not quite as the real one. But a math-

ematician is content if the tree growth strong and beautiful regardless of the non-mathematical origin of the seed.

Can one predict which seed will grow to a tree and which is a dead grain of dust, not a seed at all? This is the first and the most formidable question facing a mathematician who is looking for a problem from science: *identify a mathematically promising problem.*

Even in the pure mathematics the viability of a seed is seen only with the hindsight. Who could divine $e^{2\pi i} = 1$ and the Riemann mapping theorem at seeing $\sqrt{-1}$ as the 90°-rotation in the plane, or realize that the "reduction" $4 \Rightarrow 3$ implemented by the (exactly!) *three* $2 + 2$ partitions of the *four* element set solves algebraic equations of degree 4 (modulo degree 3), points toward the Yang-Mills equations and would grow in a proper soil to the Donaldson theory?

We present below two ideas coming from biology:

Menedel's multilinear dynamics that has undergone a significant mathematical developement but on the essentially 19th century soil –"mathematics of the multiplication-table type" citing Godfrey Harold Hardy,

and

Sturtevant's structure recognition paradigm that has not been absorbed yet by the present day mathematics.

Mendel's Model of Heredity. Imagine some species of flowers that come in two possible *pure* colors, say white and red but nothing else, *no pink*, no matter how you interbreed them. Moreover,

•*certain red couples have white as well as red children;*

•• *sometimes a red couple has the descendants in all generations being exclusively red;*

••• *the descendants of white parental flowers are always white.*

A possible mechanism was suggested by Gregor Mendel in his paper "Versuche über Pflanzen-Hybriden" that appeared in *Verhandlungen des naturforschenden Vereines, Abhandlungen, Brunn 4, pp. 3-47 (1866)* (seven years after Charles Darwin's Origin of Species) where Mendel writes in the introduction: "The striking regularity with which the same hybrid forms always reappeared whenever fertilization took place between the same species induced further experiments to be undertaken, the object of which was to follow up the developments of the hybrids in their progeny".[1]

He postulated that each *pure* phenotypic feature F of an organism, such as the above color, is determined by what is now-a-days called the *gene* occupying a specific *locus* responsible for F. Mendel hypothized that each gene was made of two "halves", one inherited from the mother and one from the father. We denote by $A = A(locus)$ the set of all such possible "halves", called *allels* ("alternative (half)genes" usually called *gametes*[2]) and think of F as a function of the gene composition, i.e. a function in two A-variables, $F = F(a, b)$, $a, b \in A$, where, after Mendel it is assumed to be *symmetric* and genes are formally written as *quadratic monomials* in the A-variables, ab's instead of (a, b)'s.

In the case of flowers, a natural candidate for A is a two point set $A = \{r, w\}$, where the presence of the r-allele in the gene ensures the production of the red

[1]The original article as well as its English translation can be found at http://www.mendelweb.org/

[2]Our choice of words is adapted to their mathematical usage and often deviates from traditional terminology accepted by geneticists.

color pigment, while the w-allele is color deficient[3]. This suggests the following values of the function[4] $F = COLOR$,

$F(ww) = white$, $F(wr) = red$, and $F(rr) = red$.

In the above example one can not discriminate between wr and rr allele composition by the apparent color, since both wr and rr flowers are red; yet, this is possible with the following rule:

$\star$ *parents with* ab *and* $a'b'$ *genes may have children of four kinds:*

aa', ab', ba' and bb'.

In particular, among (sufficiently many) children of two rw parent flowers there will be whites as well as reds but rw and rr parents will have all their children red. These two may have, however, white grandchildren (with two rw parents) but *all* descendants of an rr couple will be red.

This distinguishes rr from rw but is not sufficiently quantitative so far. Mathematics truly enters with the following

Mendel's Rule. *The above four outcomes* aa', ab', ba' *and* bb' *, are equiprobable.*

For example, if both father and mother are rw then rr- and ww-children come with probabilities both equal to $1/4$, while the probability of an rw-child is $1/2$.

Mendel supported his ideas by cultivating and testing (properly interbred) ≈ 30000 pea plants and provided, for example, the data confirming the expected 3:1 proportion of the numbers, $|reds|/|whites|$ for the children of two rw parents, where $|...|$ denote the cardinalities of the respective sets.

Remarks. (a) The above "halves" have nothing to do with the two strands of DNA but rather with *diploidy* of certain organisms, e.g. human and pea plants, who have two sets of chromosomes inherited from the two parents.

(b) There is a relatively small group of genes, namely the genes positioned on the *sex chromosomes*[5],where the symmetry between the parental genes breaks down. However, following a tradition common in biology, we make unconditional statements allowing exceptional cases.

(c) Mendel's postulate expresses a common "equalizing idea" in mathematical modelling: *two similar constants are assumed equal unless there is information to the contrary.*

(d) The data provided by Mendel, such as the sharpness of the above 3:1 proportion, looked so good, that in 1936, R.A. Fisher, one of the founders of the population genetics, concluded, on the basis of a χ^2-analysis, that Gregor Mendel had falsified his data.[6]

[3]Such a pigment is usually synthesized in some cells of a flower plant by concerted activity of several proteins P along a specific *metabolic pathway* where one of them, say $P_\star$, may be crucial for the pigment production. Such a $P_\star$, as we know it now-a-days, is *coded* by the corresponding gene – a specific segment on the (very long) DNA molecule making a *chromosome*. Every diploid cell contains two sets of chromosomes and thus two genes coding for $P_\star$, where one of the two may code for a protein $P'_\star$ that is slightly different from $P_\star$ and does not work properly in the pigment synthesis.

[4]The function $F = F(a, b)$ can be often reduced to a function in one variable by something like $F(a, b) = max(f(a), f(b))$. In the case of flowers the production of a pigment by $P_\star$ coded on a single chromosome is sufficient for the ample color.

[5]http://en.wikipedia.org/wiki/Chromosome
http://biology.about.com/library/weekly/aa091103a.htm

[6]See http://www.mcn.org/c/irapilgrim/men05.html
for a criticism of Fisher's view.

(e) Our exposition is too short for what it stands for– one of the most intellectually innovative step ever made in biology and arguably in the all of science. Curt Stern stated in 1966: "Gregor Mendel's short treatise, 'Experiments on Plant Hybrids' is one of the triumphs of the human mind. It does not simply announce the discovery of important facts by new methods of observation. Rather, in an act of highest creativity, it presents these facts in a conceptual scheme which gives them general meaning. Mendel's paper is not solely a historical document. It remains alive as a supreme example of scientific experimentation and profound penetration of data".[7]

Unlike the ready acceptance/rejection of the Darwinian Origin of Species by scientists and laymen, it took several generations of biologists to absorb Mendel's ideas and to reconcile the (suitably modified) Darwinian concept of selection and evolution with Mendelian genetics.

Hardy-Weinberg Principle for Allele Distributions. A *distribution* on a set X or an X-*distribution* is an assignment of a real (often positive integer) weight n_x to each $x \in X$ where only finitely many among n_x are different from zero. (In what follows the relevant sets are finite anyway). We write such distributions as formal (finite) linear combinations $\mathbf{x} = \sum_x n_x x$ and regard them as linear forms, i.e. polynomials of degree 1 in the x-variables. We denote by $\mathbb{R}(X) \supset X$ the set of all distributions and observe that every map $X \to Y$ extends to a linear map $\mathbb{R}(X) \to \mathbb{R}(Y)$, where the correspondence $X \rightsquigarrow \mathbb{R}(X)$ is a covariant functor. Moreover, every map from X to a linear space R (e.g. $X \to \mathbb{R}(Y)$) uniquely extends to a linear map $\mathbb{R}(X) \to R$.

$\langle\rangle$-Function and the group $\mathcal{G}_{\langle\rangle}$. The space $\mathbb{R}(X)$ comes along with a natural linear function on it (invariant under permutations of x's) that is $\mathbf{x} = \sum_x n_x x \longmapsto \langle\mathbf{x}\rangle = \langle\mathbf{x}, \mathbf{1}\rangle =_{def} \sum_x n_x$, where $\mathbf{1}$ denotes the distribution assigning weights 1 to all x's. We shall see presently that the standard computations in formal genetics are invariant under the group $G_{\langle\rangle} = GL(\mathbb{R}(A), \langle\rangle)$ of linear transformations of $\mathbb{R}(A)$ which preserve $\langle\rangle$ and even under a *larger* symmetry group $\mathcal{G}$ defined in the next section.

On measure and probability. Distributions with positive weights are naturally identified with finite measures on X and if, moreover $\langle\mathbf{x}\rangle = \sum_x n_x = 1$, we usually write p_x instead of n_x and regard sums $\sum_x p_x x$ as probability measures (distributions) on X, where our notations follow the algebraic rather than the analytic tradition.

Gene Distributions. These are distributions on the Cartesian square $A \times A$ that are symmetric under the involution $(a, b) \longmapsto (b, a)$. We represent them by polynomials in a-variables, $\sum_{a,b} n_{a,b} ab$, where, recall, monomials represent genes at a given locus.

Allele Content Map For each gene $g = ab$ we denote by $\mathbf{a}_g$ its *allele content* defined by $\mathbf{a}_g = a + b$ that is a distribution on A. The resulting *content map* $G \longrightarrow \mathbb{R}(A)$, for $g \longmapsto \mathbf{a}_g$ (linearly) extends to its linear counterpart $\mathbb{R}(G) \longrightarrow \mathbb{R}(A)$ for

$$\mathbf{g} = \sum_{a,b} n_{a,b} ab \longmapsto \mathbf{ag} = \sum_{a,b} n_{a,b}(a + b)$$

[7] Compare http://www.weloennig.de/mendel02.htm
http://www.library.adelaide.edu.au/digitised/fisher/144.pdf

.

Since

$$\sum_{a,b} n_{a,b}(a+b) = \sum_a (\sum_b n_{a,b})a + \sum_b (\sum_a n_{a,b})b,$$

the above linearized allele content map $\mathbb{R}(G) \longrightarrow \mathbb{R}(A)$ has the following

Additivity Property. *If (the quadratic polynomial representing)* $\boldsymbol{g}$ *splits into product of two linear forms (polynomials),*

$$\boldsymbol{g} = \boldsymbol{ab} = \sum_a n_a a \sum_b n_b b = \sum_{a,b} n_a n_b ab$$

, then $\boldsymbol{a_g}$ *equals a linear combination of* $\boldsymbol{a}$ *and* $\boldsymbol{b}$*, namely*

$$\boldsymbol{a_g} = \langle \boldsymbol{b} \rangle \boldsymbol{a} + \langle \boldsymbol{a} \rangle \boldsymbol{b}.$$

In particular, if $\boldsymbol{a}$ *and* $\boldsymbol{b}$ *are probability distributions then* $\boldsymbol{a_g} = \boldsymbol{a} + \boldsymbol{b}$*; moreover, if* $\boldsymbol{a} = \boldsymbol{b}$*, i.e.* $\boldsymbol{g} = \boldsymbol{a}^2$*, then* $\boldsymbol{a_g} = 2\boldsymbol{a}$*.*

Corollary: $\mathcal{G}_{\langle\rangle}$-Invariance. *The linearized content map* $\mathbb{R}(G) \longrightarrow \mathbb{R}(A)$ *is equivariant under the natural action of the group* $\mathcal{G}_{\langle\rangle} = GL(\mathbb{R}(A), \langle\rangle)$ *on the space* $\mathbb{R}(G)$ *identified with the tensorial symmetric square of* $\mathbb{R}(A)$*.*

*Remark.*It is easy to see that two $\mathcal{G}_{\langle\rangle}$-equivariant maps differ by a multiplicative constant and so the equivariance property *uniquely*, up to a scalar multiple characterizes the linearized content map.

Algebraic Formulation of Mendel's Rule. *The distribution of genes of the children of a* g*- and a* g'*-parent equals the product of their allele contents,* $\boldsymbol{g}_{children(g,g')} = \boldsymbol{a}_g \boldsymbol{a}_{g'}$*.*

In fact, if $g = ab$ and $g' = a'b'$, then $\mathbf{a}_g \mathbf{a}_{g'} = (a+b)(a'+b') = aa' + ab' + ba' + bb'$ with the agreement with the equiprobability formulation of his rule.

Remark on normalization. If one insists on *probability* distributions (that we do not always do) one has to normalize the above by taking $\frac{1}{4} a_g a_{g'} = \frac{1}{4} aa' + \frac{1}{4} ab' + \frac{1}{4} ba' + \frac{1}{4} bb'$. Or equivalently, one had to divide the content map by 2 in order to make it $\langle\rangle$-preserving.

"Next Generation" Map for Random Mating. Let X be a populations of organisms of some species and g_x denote the gene of $x \in X$ at a given locus. Then the distribution of genes in X is $\mathbf{g}_X = \sum_x g_x = \sum_g n_g g$, where n_g denotes the number of individuals (organisms) in X carrying gene g, i.e. with $g_x = g$ and where, observe, the sun $n = \sum_g n_g$ equals the cardinality $\#X$. If we pick *on random* some x from X its gene is represented by the probability distribution $\mathbf{g} = \sum_g p_g g$ for $p_g = \frac{1}{n} n_g$. If we take *independetly* another "random individual", possibly from another population X' with gene distribution $\mathbf{g}' = \sum_{g'} p_{g'} g'$ then the genes of their children will be distributed by *the random next generation rule*:

$$\mathbf{g}_{children(\mathbf{g},\mathbf{g}')} = \sum_{g,g'} p_g p_{g'} \mathbf{g}_{children(g,g')}.$$

Randomness and Symmetry. The "randomness" amounts in the present context to the *symmetry* of our choice of an x with respect to the permutation group acting on X: a "random individual" $\mathbf{x}$ is represented by the probability

distribution on X assigning *equal* weights($=\frac{1}{n}$) to all $x \in X$, where the corresponding distribution on genes, $\mathbf{g}_X = \mathbf{g}_{\mathbf{X}} = \sum_x g_x = \sum_g n_g g$ equals the value at $\mathbf{x}$ of the linear extension of the map $x \longmapsto g_x$ to a map $\mathbb{R}(X) \longrightarrow \mathbb{R}(G)$. Random $\mathbf{x}'$ has a similar meaning while "independence" of the two choices, encoded by *bilinearity* (in p_g and $p_{g'}$) of the above formula for children, reflects the symmetry of (probability of) mating with respect to the permutation group acting on $X \times X'$.

Mendel's rule for random mating. *The genes of children of independently chosen random parents are distributed by the same rule as those for individuals, namely they abide the following*

Allele Product formula:

$$\boldsymbol{g}_{children(\boldsymbol{g},\boldsymbol{g}')} = \boldsymbol{a}_{\boldsymbol{g}}\boldsymbol{a}_{\boldsymbol{g}'}.$$

In particular, if the parents have equal allele distributions, say $\boldsymbol{a}$*, then the next generation gene distribution equals the square of this allele distribution,*

$$\boldsymbol{g}_{children} = \boldsymbol{a}^2.$$

Proof. Substitute $\mathbf{g}_{children(g,g')}$ in the "random next generation" rule with $\mathbf{a}_g\mathbf{a}_{g'}$ in accordance to (individual) Mendel's rule and decompose $\sum_{g,g'} p_g p_{g'} \mathbf{a}_g \mathbf{a}_{g'}$ into the product $(\sum_g p_g \mathbf{a}_g)(\sum_{g'} p_{g'} \mathbf{a}_{g'})$.

Corollary A. *The gene distribution of "random children" depends only on the allele distributions of their parents rather than on the parental gene distribution in the population.*

Since allele distributions depend on $k = \#A$ parameters while gene distributions depend on $\#G = \frac{k(k+1)}{2}$ parameters, the corollary reduces the dimension from $\frac{k(k+1)}{2}$ to k.

Corollary B. *The allele distribution of children is expressed by these of their random parents* $\boldsymbol{x}$ *and* $\boldsymbol{x}'$ *according to the*

Allele Addition Formula (Leibniz rule): $\mathbf{a}_{children} = \langle \mathbf{a}_{\mathbf{X}'} \rangle \mathbf{a}_{\mathbf{X}} + \langle \mathbf{a}_{\mathbf{X}} \rangle \mathbf{a}_{\mathbf{X}'}$.

In particular, if both parents have equal gene (or just allele) distributions then the *normalized* allele distribution for the children equal those of the parents (where the normalization amounts to scaling the outcome of the above formula by $\frac{1}{2}$ thus making it a *probability* distribution). In other words,

the normalized "next generation" map on the space of allele probability distributions equals the identity, i.e. the distribution of alleles in a population does not change under random mating.

Remarks.(a) If one of the parents serves as a male and the other one is female, their gene distributions can be, a priori, different but if we deal, for example, with self pollinating plants then we may assume that the parents are taken from the same population X and thus have equal gene and allele distributions.

(b)The *stability of the allele distribution under random mating* can be also derived from the $\mathcal{G}_{\langle\rangle}$-equivariance of the *random mating map* $\mathbb{R}(A) \longrightarrow \mathbb{R}(A)$ that assigns to each parental allele distribution that of the children. This map (when normalized in order to preserve probability distributions) fixes the monomials (that make the basis) in the space $\mathbb{R}(A)$ and hence fixes all points since the

orbit of each monomial (basis vector) under the action of $\mathcal{G}_{\langle\rangle} = GL(\mathbb{R}(A), \langle\rangle)$ is (Zariski) dense.

Summary: (Castle)-Hardy-Weinberg Equilibrium Principle. *Assume that both (populations of) parents have equal gene distributions, say* $\boldsymbol{g}_0$*, and denote by* $\boldsymbol{g}_1, \boldsymbol{g}_2, ...$ *the gene distributions of their random children, grand children, etc., that are normalized to probability distributions. Then* $\boldsymbol{g}_1 = \boldsymbol{g}_2 = ...$ *In fact,* $\boldsymbol{g}_i = (\frac{1}{2}\boldsymbol{a}\boldsymbol{g}_{i-1})^2$ *where* $\boldsymbol{a}\boldsymbol{g}_i = \boldsymbol{a}\boldsymbol{g}_{i-1}$ *for* $i \geq 1$*. If, moreover,* $\boldsymbol{g}_0$ *equals the square of a linear polynomial then also* $\boldsymbol{g}_0 = \boldsymbol{g}_1$ *and the square condition is necessary as well as sufficient for this equality.*

Remarks. (a)The proof of the equilibrium property appears in Hardy's one page letter to the editor: *"Mendelian proportions in a mixed population". Science 28: 49a50 (1908)* , where Hardy, who deals with the symmetric 2×2 -matrices, (proportionly) represented by the numbers $p : 2q : r$, writes

"...suppose that the numbers are fairly large, so that mating may be regarded as random, that the sexes are evenly distributed among the three varieties, and that all are equally fertile. A little mathematics of the multiplication-table type is enough to show that in the next generation the numbers will be as $(p+q)^2 : 2(p+q)(q+r) : (q+r)^2$, or as $p_1 : 2q_1 : r_1$, say.

The interesting question is in what circumstances will this distribution be the same as that in the generation before? It is easy to see that the condition for this is $q^2 = pr$. And since $q_1^2 = p_1 r_1$, whatever the values of p, q, and r may be, the distribution will in any case continue unchanged after the second generation."

These nine lines had cleared up the confusion of Hardy's contemporaries on the implications of Mendel's theory[8] and, ironically, brought Hardy the fame exceeding that of his as a pure mathematician. (Google's ratio ("Hardy theorem" + "Hardy-Littlewood theorem"):("Hardy-Weinberg law ") is about 1 : 30.)

(b) Instead of normalizing, one could *projectivize*, i.e. factor away normalizing scalars, and then express the Equilibrium Principle by saying that the projectivized "next generation" map R is a *retraction* or an *idempotent*, i.e. $R \circ R = R$; our R retracts the projective space $P\mathbb{R}(G)$ onto the subspace $P\mathbb{R}(A) \subset P\mathbb{R}(G)$, where the projectivized allele space $P\mathbb{R}(A)$ is embedded to $P\mathbb{R}(G)$ via the so called *Segre-Veronese map* $\mathbf{a} \longmapsto \mathbf{g} = \mathbf{a}^2$ and where the retraction $P\mathbb{R}(G) \longrightarrow P\mathbb{R}(A)$ is a rational (rather than regular) map whose all fibers are projective subspaces in $P\mathbb{R}(G)$.

(c) All of the above trivially generalizes to *d-ploid* organisms having $d \geq 2$ copies of each chromosome where the corresponding gene distribution space is represented by the space of homogeneous polynomials $\mathbf{g}$ in a-variables of degree d. Here the normalized next generation map can be conveniently described with the $(d-1)$th power of the operator $\delta\mathbf{f} = (deg\mathbf{f})^{-1}\partial_{\mathbf{1}} f$ acting on polynomials $\mathbf{f}$ of all degrees, where $\partial_{\mathbf{1}}$ denotes differentiation along the vector $\mathbf{1} = (1, 1, ..., 1)$; namely, $\mathbf{g}_{children} = (\delta^{d-1}\mathbf{g}_{parents})^d$ and the Equilibrium Principle amounts to

[8]see http://en.wikipedia.org/wiki/Hardy-Weinberg. The point made by Hardy, as I see it, was not the "multiplication table" but rather identifying "random" (mating) "evenly distributed" and "equally" (fertile) as mathematical concepts. This, most probably, was obvious to Mendel by 1866 and if, fantasizing, Riemann (who died in 1866) had become acquainted with Mendel's explanation of "the striking regularity with which the same hybrid forms always reappeared" he would have been be amazed and delighted, unlike the biologists of that time who dismissed Mendel's results as "non-interesting".

the easily verifiable identity: $(\delta^{d-1}(\delta^{d-1}\mathbf{g})^d)^d = (\mathbf{g}(\mathbf{1})^{d^2-d}(\delta^{d-1}\mathbf{g})^d$, that is valid for all homogeneous polynomials $\mathbf{g}$ of degree d.

(d) The "random mating" assumption is rather stringent and is hardly ever observed in its pure form. In reality, there may be some *selection* mechanism at work that distorts Menedel's equiprobability rules and makes the next generation map more complicated. For example, a breeder may systematically eliminate white flowers from the breeding pool. In a case like that one still has a homogeneous quadratic "next generation" self map on the space $\mathbb{R}(G)$ (and/or on $P\mathbb{R}(G)$) which often reduces to a similar kind of map on $\mathbb{R}(A)$. Such map does not have to stabilize at a finite step but it may asymptotically converge to some stable attractive fixed point(s), where the likely candidates for such points in simple models are the vertices of the unit simplex $\Delta \subset \mathbb{R}(A)$ spanned by the monomials and/or the center of the simplex. (see [2])

Also one may take into account restrictions imposed by positions occupied by organisms, and hence by their genes and/or alleles, in a physical space: individuals mate preferably with their neighbors and children remain in the vicinity of the parents, where an extremal case is that of exclusively self pollinating plants.

It is not hard for a mathematician to come up at this point with a variety of specific aesthetically attractive models (such as *Kolmogorov-Petrovskii-Piskunov equation*) but it is hard, even for a biologist, to pinpoint a biologically feasible one.

(e) The Mendel rules are similar to the *law of mass action* in the *ideal chemical kinetics*[9] (the first approximation to the "true chemistry") formulated by Cato Maximilian Guldberg and Peter Waage[10]. This law says that the rate of transformations of compounds A_i to B is proportional to the *product* of the concentrations of A_i, since the probability of the thermally moving molecules of A_i in a pot coming sufficiently close together in order to participate in a reaction is proportional to the product of their concentrations. (If the production of a molecule of B needs k_i molecules of A_i the concentration of A_i enters in the k_ith power.) This leads to a multilinear system S of ODE on the concentrations of the compounds, that is a polynomial vector field on the Euclidean n-simplex $\Delta_n \subset \mathbb{R}^{n+1}$ of the (normalized) concentrations. (If the production of a molecule of B needs k molecules of some A_i the concentration of A_i enters in the kth power.)

The fundamental mathematical problems arising in chemistry that (similarly to those in the Mendelian genetics) can not be expressed in the language of the smooth dynamical systems, i.e. in terms of invariants of transformations up to conjugation in the group of *all* diffeomorphisms (homeomorphisms). There is more to the structure in S than mere $Diff$: the dimension n is not "just a number" but a combinatorial object implemented by the vertex set V of a (weighted) graph with an additional structure reflecting the hierarchy of the rates of different reactions. Operation normally performed by a chemists (introducing a catalyzer, removing a product, etc.) "naturally" correspond to transformations/degenerations of graphs that are, in turn, functorially reflected by the dynamics, where the objects corresponding to "degeneration" are asymptotic limits of S that are not dynamical systems in the ordinary sense.

[9] see http://www.sussex.ac.uk/chemistry/documents/rates.pdf for a historical overview.

[10] see http://chimie.scola.ac-paris.fr/sitedechimie/hist_chi/text_origin/guldberg_waage/Concerning-Affinity.htm for the English translation of the original 1864 Norwegian presentation.

Besides, while a mathematician strives on the dynamical subleties of apparantly simple systems (e.g. diffeomorphisms of the circle), a scientists seeks the islands of simplicity in the ocian of impenetrable complexity of the "real world" dynamics.

In the case of the *linear* systems, the separation of the reaction rates allows (see [3]) a reduction of the continues dynamics to the combinatorial one on V (with the numerics associated the metric/measure invariants of the singularities of the discriminant variety in the space of linear operators that has a flavor of the "tropical reduction" in the algebraic geometry).

There is no conceptual mathematical framework yet for non-linear systems (despite a huge number of particular systems that has been analyzed) but there are some combinatorial criteria for a "simple/robust" behavior of S that are (believed to be) frequently fulfilled in the metabolic pathways , for instance [1].

Recombination. The next generation map $\mathbf{g} \longmapsto (\frac{1}{2}\mathbf{a}\mathbf{g})^2$ can be defined on the space $\mathbb{R}(A \times A) = \mathbb{R}(A) \bigotimes \mathbb{R}(A)$ of *all*, not only symmetric distributions on $A \times A$ represented by $k \times k$ matrices for k denoting the number of the alleles (i.e. the cardinality of A) as follows,

substitute each (i, j)-entry in such a matrix by the product of the sum of entries in the i- row by the sum of entries in the j-column.

Here the proof of the equilibrium property reduces to a tautology as follows.

Consider two linear spaces $\mathbf{A}$ and $\mathbf{B}$ with distinguished non-zero linear functions on them both denoted $\langle\rangle$. The tensor product $\mathbf{C} = \mathbf{A} \bigotimes \mathbf{B}$ of such spaces is given $\langle\rangle$ by linearly extending $\langle \mathbf{c} \rangle = \langle \mathbf{a} \rangle \langle \mathbf{b} \rangle$ from monomials $\mathbf{c} = \mathbf{a} \bigotimes \mathbf{b}$ to all of $\mathbf{C}$. There are two $\langle\rangle$-natural linear maps from $\mathbf{C} = \mathbf{A} \bigotimes \mathbf{B}$ to the tensorial components: $E_{\mathbf{A}} : \mathbf{C} \longrightarrow \mathbf{A}$ is defined as the linear extension of the (bilinear) map $\mathbf{a} \bigotimes \mathbf{b} \longmapsto \langle b \rangle \mathbf{a}$ and similarly one defines the map $E_{\mathbf{B}}$ to $\mathbf{B}$. ($E_{\mathbf{A}}$ and $E_{\mathbf{B}}$ correspond to summations of rows and column in matrices). With these two we define the "next generation map" E from $\mathbf{C}$ to itself by $E(\mathbf{c}) = E_{\mathbf{A}}(\mathbf{c}) \bigotimes E_{\mathbf{B}}(\mathbf{c})$,where the equilibrium property reads:

$E \circ E(\boldsymbol{c}) = \langle \boldsymbol{c} \rangle E(\boldsymbol{c})$; thus E is an idempotent (i.e. $E \circ E = E$) on the subset (hyperplane) in $\mathbf{C}$ of vectors normalized by $\langle \boldsymbol{c} \rangle = 1$.

Indeed, E($\mathbf{c}$) is a monomial, and *each* monomial, say $\mathbf{c}' = \mathbf{a}' \bigotimes \mathbf{b}'$ goes under E to $\langle \mathbf{b}' \rangle \mathbf{a}' \bigotimes \langle \mathbf{a}' \rangle \mathbf{b}' = \langle \mathbf{c}' \rangle \mathbf{c}'$.

Let us generalize the above to multiple tensor products of $\langle\rangle$-spaces, $\bigotimes_{l \in L} \mathbf{A}_l$ for an arbitrary finite set L. Such a product can be seen as a subspace in the polynomial algebra $\mathbf{A}^* = \mathbf{A}^*(X)$ on the Euclidian space X that is the sum (Cartesian product) $\bigoplus_{l \in L} X_l$ of the linear spaces X_l dual to $\mathbf{A}_l$: the product $\bigotimes_{l \in L} \mathbf{A}_l$ is identified with the set of homogeneous polynomials of degree 1 in each x_l-variable where $\langle \mathbf{a} \rangle$ is represented by the value $\mathbf{a}(x_0)$ at some vector $x_0 \in X$. Since one can go from one vector to another by a parallel translation of X and translations induce automorphisms of the algebra $\mathbf{A}^*(X)$, the choice of x_0 makes no difference; in what follows, instead of taking $x_0 = \mathbf{1} = (1, 1, ..., 1)$ as we did for distribution spaces in the previous section, we save notation by taking $x_0 = 0$ in X.

With each subset $K \subset L$ we associate the coordinate projection P_K from X to the coordinate plane $X_K = \bigoplus_{l \in K} X_l \subset X$ and denote by $E_K = P_K^*$ the induced endomorphism of the algebra $\mathbf{A}^*$. (In simple words, applying E_K

to an $\mathbf{a}(x_l)$ amounts to equating all x_l in $\mathbf{a}$ with $l \in L - K$ to zero.) Since P_K are *commuting idempotents* so are E_K for all $K \subset L$, where E associated to the empty set sends $\mathbf{A}^*$ to the constants. Given a collection $\mathcal{K}$ of subsets $K \subset L$ we define $E_\mathcal{K}$ as the (polynomial) product of E_K for all $K \in \mathcal{K}$, i.e. $E_\mathcal{K}(\mathbf{a}) = \prod_{K\in\mathcal{K}} E_K(\mathbf{a})$. Since the multiplicative semigroup of polynomials is commutative and the maps E_K are endomorphisms, the transformations $E_\mathcal{K}$ are *multiplicative* endomorphisms of $\mathbf{A}^*$ (but not additive ones for more than one K in $\mathcal{K}$). Since all E_K commute, so do $E_\mathcal{K}$ and the composition of $E_\mathcal{K}$'s is expressible in terms of intersections of the underlying subsets $K \subset L$ by the simple rule: $E_\mathcal{K} \circ E_{\mathcal{K}'} = \prod_{K\in\mathcal{K},K'\in\mathcal{K}'} E_{K\cap K'}$, that follows from the similar rule for the composition of the maps P_K's.

Equilibrating Maps. If $\mathcal{K}$ is made of d *non-intersecting* non-empty subsets, e.g. $\mathcal{K}$ is a partition of L into d subsets, then $E = E_\mathcal{K}$ is called an *equilibrating* map of *degree* d. Equilibrating maps obviously satisfy:

(A) *Composition property.* A composition of equilibrating maps of degree d and d' is an equilibrating map of degree dd' with the following "self-composition" rule: $E \circ E(\mathbf{a}) = \mathbf{a}(0)^{d^2-d} E(\mathbf{a})$, where the exponent corresponds to the presence of d^2-d *empty* intersections between different subsets $K_1, ..., K_d$ in L underlying E.

(B) *Polynomiality.* Equilibrating maps preserves subspaces $\mathbf{A}^{\leq k} \subset \mathbf{A}^*$ of polynomials of degree $\leq k$ in each variable. Thus $\mathbf{A}^*$ is representable as a union of *finite dimensional* E*-invariant* subspaces and if $\mathcal{K}$ is made of d subsets $K \subset L$ then the corresponding equilibrating map is a polynomial map of degree d on each linear space $\mathbf{A}^{\leq k}$.

(C) *Linearizability.* One can regard $\mathbf{A}^{\leq k}$ as the algebra of k*-truncated polynomials* that is a quotient of (rather than a subspace in) $\mathbf{A}^*$ obtained by adding the relations $x_l^{k+1} = 0$ to $\mathbf{A}^*$. The maps $E_\mathcal{K}$ (not only equilibrating ones) act on this algebra as multiplicative endomorphisms; they can be "simultaneously linearized" with the exponential map, $exp(\mathbf{a}) = 1 + \mathbf{a} + \frac{1}{2}\mathbf{a}^2 + \frac{1}{6}\mathbf{a}^3 + ...$, that isomorphically maps the *additive group* of k-truncated polynomials to the *multiplicative group* of k-truncated polynomials satisfying $\mathbf{a}(0) > 0$.

(D) *Retraction to Veronese.* It follows from (A) (and also from (C)) that each equilibrating map $E = E_\mathcal{K}$, $\mathcal{K} = (K_1, K_2, ..., K_d)$, *retracts* the *normalizing hyperplane* $\mathbf{A}^\times = \mathbf{A}^\times(X) \subset \mathbf{A}^*$ defined by $\mathbf{a}(0) = 1$ to the *Veronese* product set $\mathbf{V} = \mathbf{V}_E = E(\mathbf{A}^\times) = \mathbf{A}_1^\times \cdot \mathbf{A}_2^\times \cdot ... \cdot \mathbf{A}_d^\times \subset \mathbf{A}^\times$ for $\mathbf{A}_i^\times = \mathbf{A}^\times(X_{K_i})$, that is the set of products of d polynomials $\mathbf{a}_i \in \mathbf{A}_i^\times$, where composition of E's corresponds to intersection of V's: $\mathbf{V}_{E\circ E'} = \mathbf{V}_E \cap \mathbf{V}_{E'}$.

The fibers $E^{-1}(\mathbf{v}) \subset \mathbf{A}^\times$ are *affine* subspaces: they are, obviously, equal the fibers of the additive counterpart to $E = E_\mathcal{K}$, that is $E_{K_1} + ... + E_{K_d}$, where $K_i \subset L$ are the constituents of $\mathcal{K} = (K_1, ..., K_d)$.

(E) $\mathcal{G}$*-equivariance.* The equilibrating maps E *commute* with the group $\mathcal{G}$ of linear transformations of X preserving the decomposition $X = \bigoplus_{l\in L} X_l$ that naturally act on polynomials. (For example, the Veronese varieties are $\mathcal{G}$-invariant.) In particular, All E commute with the *scaling transformation* Λ corresponding to $x \mapsto \lambda x$ in X which fixes constant polynomials, e.g. $\mathbf{1} \in \mathbf{A}^\times$, and has other eigenvalues equal $\lambda, \lambda^2, \lambda^3$, etc. Thus, for $\lambda > 1$, the transformation Λ *expands* $\mathbf{A}^\times$ with the fixed point $\mathbf{1}$ and so *global* properties of maps commuting with Λ, e.g. of equilibrating maps and linear combinations of these, can be derived from the corresponding *local* ones at the fixed point $\mathbf{1}$ of Λ by transporting all points close to $\mathbf{1}$ by applying Λ^{-N} with large $N \to \infty$.

Remark on Λ-equivariant maps. Let A be a linear space (e.g. $\mathbf{A}^\times \bigcap \mathbf{A}^{\leq k}$ with the constant polynomial $\mathbf{1}$ taken for the origin) with a linear transformation Λ, where A splits into n eigenspaces of Λ with the corresponding eigenvalues $\lambda, \lambda^2, ..., \lambda^n$, where λ is *not* a root of unity, e.g. $\lambda > 1$. It is easy to see that every smooth transformation F of A commuting with Λ is a polynomial map of degree at most n; a transformation F is invertible (necessarily by a *polynomial* transformation) if and only if its differential $D_0(F)$ at 0 is invertible; transformations F with $D_0(F) = 1$ make a nilpotent Lie group. For example, *all* iterates F^j are polynomials of degrees bounded by *the same* n that, non-surprisingly, admit explicit (albeit complicated) expression in terms of n. (See [5]).

Robbins-Geiringer Convergence Property. Consider a convex combination $F = c_1E_1 + c_2E_2 + ... + c_mE_m$ of equilibrating maps $E_1, E_2, ..., E_m$ restricted to $\mathbf{A}^\times$. Since $c_1 + c_2 + ... + c_n = 1$ and since all E_i fix the Veronese variety $\mathbf{V} = E(\mathbf{A}^\times) = \bigcap_i \mathbf{V}_i$ of the composition $E = E_F = E_1 \circ E_2 \circ ... \circ E_m$, so does F and for the same reason F sends each (affine!) fiber $E^{-1}(\mathbf{v})$ into itself.

The differentials $D_{\mathbf{1}}$ of E_i on $\mathbf{A}^\times$ at $\mathbf{1}$ have all their eigenvalues≤ 1 where the equalities are achieved on the vectors tangent to the corresponding Veronese varieties $\mathbf{V}_i = E_i(\mathbf{A}^\times)$, because E_i are smooth *retractions* to $\mathbf{V}_i$ (and where the eigenvalues equal 0 tangentially to their respective fibers).

The differential of F equals the convex combination of those of E_i; if we assume all $c_i > 0$, we conclude that all eigenvalues of the differential $D(F)$ on $\mathbf{V}$ on the tangent vectors transversal to $\mathbf{V}$ are< 1, since the tangent space to $\mathbf{V}$ equals the intersection of those to $\mathbf{V}_i$. (Tangentially to $\mathbf{V}$ the eigenvalues of $D(F)$ equal 1 since $\mathbf{V}$ is fixed under F.) In other words, the differential $D(F)$ *strictly contracts* the tangent vectors at $\mathbf{V}$ that are transversal to $\mathbf{V}$. It follows that F also contracts some neighborhood $\mathbf{U} \subset \mathbf{A}^\times$ of $\mathbf{V}$; therefore, each point $\mathbf{v} \in \mathbf{U}$ exponentially fast approaches $\mathbf{V}$ under iterates of F. In fact, the F-orbit of $\mathbf{v}$ converges to $E(\mathbf{v}) \in \mathbf{V}$ since F preserves the fibers of E.

This local property obviously globalizes with the expanding transformation Λ from (E) and shows that:

If all c_i are strictly positive, then the iterates $F^1 = F$, $F^2 = F \circ F^1, ..., F^j = F \circ F^{j-1}, ...$ on $\boldsymbol{A}^\times$ converge to the equilibrating map $E = E_F : \boldsymbol{A}^\times \longrightarrow \mathbf{V} \subset \boldsymbol{A}^\times$, where the convergence is uniform and exponentially fast on the compact subsets in $\boldsymbol{A}^\times \bigcap \boldsymbol{A}^{\leq k}$ for all $k = 1, 2, ...$

Remark. This conclusion remains valid if we replace the projections P_K by transformations $P_{K,\varepsilon} =_{def} x \mapsto (1 - \varepsilon)x + \varepsilon P_K(x)$, $0 < \varepsilon < 1$, (that make a one parameter semigroup converging to P_K for $\varepsilon \to 1$) and construct F with the corresponding endomorphisms $E_{K,\varepsilon}$'s for some $\varepsilon = \varepsilon_K > 0$ instead of plain E_K's. If one takes infinitesimally small ε's, one obtains a vector field on $\mathbf{A}^\times$, represented by a system $\mathcal{D}$ of non-linear differential equations, whose solutions F^t, $t \in \mathbb{R}_+$, describe a time-continuous version of the above F^j, $j = 0, 1, 2, ...$, where F^t, unlike F^j, are multiplicative endomorphisms on polynomials. These can be linearized with the map *exp* from the above (C) and thus one obtains an "explicit" solution of $\mathcal{D}$ in terms of elementary functions.

Crossover and Recombination. Let us return to genes and their alleles , now at several loci making a set L, i.e. instead of a single set of alleles as in the previous section we consider sets A_l, $l \in L$. The L-collections of

alleles, $a = (a_l)_{l\in L}$, i.e. points in the Cartesian product $A = \times_{l\in L} A_l$, are called *gametes*; their symmetric pairs written as monomials ab are called (partial) *genomes* or, more traditionally, *zygotes* .

The set G of genomes (zygotes) is acted upon by (commuting) involutions interchanging a_l with b_l in the monomials with $l \in K$ for all possible $K \subset L$. These involutions make the Abelian group $\Gamma = \mathbb{Z}_2^L$, that is the set of $\mathbb{Z}_2$-valued functions on L, naturally acting on G, where, observe, the diagonal involution *simultaneously* interchangings all a_l's with all b_l's acts trivially on G since $ab = ba$. There is a one-to-one correspondence between the involutions γ and partitions of L into pairs of subsets: the first one is $K_0 = fix(\gamma)$, consisting of those $l \in L$ where a_l and b_l are *not* interchanged by γ and $K_1 = supp(\gamma)$ is where a_l and b_l *are* interchanged.

A *crossover* is an arbitrary involution from Γ acting on G. Every crossover linearly acts on the space $\mathbf{G}$ of *genome distributions* that is the symmetric tensorial square of the space of gamete distributions, $\mathbf{G} = \mathbf{A}^2$ for $\mathbf{A} = \bigotimes_{l\in L} \mathbf{A}_l$ and for $\mathbf{A}_l$ denoting the space of allele distributions at the locus $l \in L$ (i.e. distributions on the set A_l).

The action of Γ linearly extends to that by distributions μ on Γ denoted $R_\mu : \mathbf{G} \longrightarrow \mathbf{G}$. If μ is a probability measure on Γ then R_μ is regarded as random crossover and is called *recombination*.

The gametes in diploid organisms, prior to (random) mating, undergo recombination (e.g. crossovers) and what they contribute to the next generation is *not* the gametes inherited from their parents but *recombined* gametes.

Let us describe what happens to distributions of gametes under the random mating composed with a crossover, where we assume that all individuals recombine according to the same γ.

We represent gamete distributions as earlier by polynomials on $X = \bigoplus_{l\in L} X_l$ and realize $\mathbf{G}$ by polynomials on the space $X \bigoplus X$ that are symmetric under the involution $(x, x') \mapsto (x', x)$. The group Γ naturally acts on $X \bigoplus X$ by interchanging x_l's with x'_l's in the $X_l \bigoplus X_l$'s, where x_l and x'_l are unmoved for $l \in K_0 = fix(\gamma)$ and they are interchanged for $l \in K_1 = supp(\gamma)$. Thus $X \bigoplus X$, temporarily denoted $X \bigoplus X'$ to keep track of who is who, splits into four spaces $X \bigoplus X' = (X_0 \bigoplus X_1) \bigoplus (X'_0 \bigoplus X'_1)$.

A gamete distribution, represented by a polynomial $\mathbf{a}(x_0, x_1)$, goes under random mating to $\mathbf{g}(x_0, x_1, x'_0, x'_1) = \mathbf{a}(x_0, x_1)\mathbf{a}(x'_0, x'_1)$. Then it recombines to $\gamma\mathbf{g} = \mathbf{g}(x_0, x'_1, x'_0, x_1) = \mathbf{a}(x_0, x'_1)\mathbf{a}(x'_0, x_1)$ whose gamete content equals $\mathbf{g}(x_0, x'_1, 0, 0) = \mathbf{g}(0, 0, x'_0, x_1) = \mathbf{a}(x_0, 0)\mathbf{a}(0, x'_1)$. This can be equally written as $\mathbf{a}(x_0, 0)\mathbf{a}(0, x_1)$, since $X = X'$ and $X_1 = X'_1$. By observing that the latter equals $E_\mathcal{K}(\mathbf{a})$ for $\mathcal{K} = (K_0, K_1)$, we see that

each crossover γ acts on $\boldsymbol{A}^*$ by the equilibrating operator $E_\gamma = E_\mathcal{K}$ associated to $\mathcal{K} = (K_0 = fix(\gamma), K_1 = supp(\gamma))$.

Consequently, a recombination $\mu = \mu(\gamma)$ acts on gamete distributions as the convex combination $F = \sum_{\gamma\in\Gamma} \mu(\gamma)E_\gamma$, and by the Convergence Property the iterates of this F converge to the equilibrating map $E = E_F$ corresponding to the partition of L into the subsets K_i defined as follows: l_1 and l_2 from L belong to two *different* subsets of the partition if and only if there is $\gamma \in \Gamma = \mathbb{Z}_2^L$, such that $\mu(\gamma) > 0$ and one of the two components γ_{l_1} and γ_{l_2} is the *trivial* involution (i.e. the identity in $\mathbb{Z}_2$) while the other one is *non-trivial*. This yields

Robbins-Geiringer Asymptotic Equilibrium Theorem. Consider a population X_0 with some gamete probability distribution $\mathbf{a} = \mathbf{a}(X_0)$, where

the corresponding allele (also probability) distributions at the loci $l \in L$ are denoted by $\mathbf{a}_l = \mathbf{a}_l(X_0)$. Observe that $\mathbf{a}_l$, are (obviously) conserved under all recombinations; they are also invariant under random mating according to Hardy-Weinberg equilibrium principle; therefore, their product (that is a certain gamete distribution generally different from $\mathbf{a}$), call it $\mathbf{a}_{equi} = \mathbf{a}_{equi}(X_0) =_{def} \prod_{l\in L} \mathbf{a}_l$ is stable under random matings and recombinations as well.

Let μ be a probability measure on the group $\Gamma = \mathbb{Z}_2^L$, such that the support of μ generates Γ. Then the gamete probability distributions $\boldsymbol{a}(X_i)$ of the populations X_0, X_1, X_2,..., X_i,... resulting from consecutive rounds of random matings and μ-recombinations, converge to $\boldsymbol{a}_{equi} = \boldsymbol{a}_{equi}(X_0)$, for $i \to \infty$.

Remarks. (a) This can be also seen by observing that the entropy (introduced by Boltzmann in 1877) of a distribution $\mathbf{a}$ increases by an $\varepsilon > 0$ at each round of the recombination-next-generation map unless $\mathbf{a}$ reaches an equilibrium (see 5.4 in [4] for the related discussion and references therein).

(b) Let $\mathbf{A}^*$ be a topological algebra and consider polynomial selfmappings

$$F = \sum_J \mu_J E^J : \mathbf{A}^* \to \mathbf{A}^*,$$

where $E^J = E_{j_1}E_{j_2}...$ are products of some endomorphisms E_{j_k} of $\mathbf{A}^*$. One can not expect much of such maps F in general as these are may be rather dense by the Weierstrass approximation theorem, but if the dynamics of the semigroup generated by E_{j_k} on the space B of the maximal ideals of $\mathbf{A}^*$, say for commutative algebras $\mathbf{A}^*$ realized by functions on B, is sufficiently simple, one may have fixed points $\mathbf{a} \in \mathbf{A}^*$ of F (equilibrium states) with controlled basins of attractions, where, moreover, these $\mathbf{a}$ maximize some (entropy) function on $\mathbf{A}^*$.

The classical example (of slightly different nature) is where $\mathbf{A}^*$ is the algebra of l_1-functions $\mathbf{a}$ on R^n under convolution and $F(\mathbf{a}(x)) = \mathbf{a}^2(\sqrt{2}x)$. The centered Gaussian *probability* measures $c\cdot e^{-Q(x)}$, where $c = (\int e^{-Q(x)})^{-1}$, are fixed under this F, they maximize the entropy among all centered measures with given second momenta and the basin of the F-attraction of the Gaussians contains all centered measures with finite second momenta.

Probably, (I could not find a reference) this remains true for more general monomial maps $F = \prod E_k$ (and, possibly, some convex combinations of these) where the endomorphims E_k are induced by linear maps $P_k : \mathbb{R}^n \to \mathbb{R}^n$ and where the needed condition on P_k (and on the basin of attraction) is seen by looking at the corresponding *linear* map on the (Tailor expansions at 0 of the) logarithms of the Fourier images $\hat{\mathbf{a}}$ of probability measures $\mathbf{a}$.

Furthermore, if X is the total space of a vector bundle with the fiberwise convolution product of measures on X, then a similar "central limit/ergodic theorem", probably, remains true for monomilas (and some polynomilas) in the endomorphisms E_k indiced by fiberwise linear self-mappings of X that satisfy suitable assumptions.

There is, yet, another setting where a map F goes from $\mathbf{A}^*$ to some tensorial power $A^{\bigotimes d}$ that is motivated by the *entropic Shannon-Loomis-Whitney-Shearer-Brascamp-Lieb inequalities* (see [4] and references therein). For example, polynomials in three groups of variables go to polynomials in six groups of variables under the map $F : \mathbf{a}(x_i, y_i, z_i) \mapsto \mathbf{a}(x_i, y_i, 0)\mathbf{a}(x_i', 0, z_i')\mathbf{a}(0, y_i'', z_i'')$. Apparently, the iterates of such maps are asymptotic in a certain sense to the

tensorial products of some equilibrium distributions that are extremal for the corresponding entropic inequalities.

Question. Is there a comprehensive theory encompassing all this?

(c) The rendition of Mendel's ideas presented in the present paper is also, in Hardy's words, "mathematics of the multiplication-table type" but now *not* with the "tables" of numbers but of something else–the rings of truncated polynomial in the above discussion. This "something" directly descends from the universality/functoriality of Mendel's (not fully formalized) model but it becomes virtually invisible once everything is reduced to mere numbers.

As it usually happens, the mathematical descendants of an idea coming from science can be seen only in the light of the abstract concepts available at the time; Mendel's inheritance rules have been studied by (applied) mathematicians in the context of the *population genetics* and *quantitative genetics* for about hundred years in the conceptual frames specific to each period (see http://en.wikipedia.org/wiki/Population genetics and references therein and also [2], [5] [6], where the "post-functorial" mathematics have not said its word yet.

Gene Linkage and Linear Arrangement of Genes. A conceptually new idea, the idea of using recombination as a device for looking *inside a cell* and seeing how genes are arranged on a chromosome came from a biologist.

In the 1913 paper *"The linear arrangement of sex-linked factors in Drosophila, as shown by their mode of association"* Alfred Sturtevant[11], long before the advent of the molecular biology and discovery of DNA, has deduced the *linearity* of the arrangement of genes on a chromosome from the statistics of simultaneous occurrences of particular morphological features in generations of suitably interbred Drosophila flies. Thus he obtained the world's first *genetic map*, i.e. he determined relative positions of certain genes on a chromosome, where he used his ideas of linearity and of *gene linkage*.

A mental picture here is as follows: genes are seen as beads on a string (chromosome), i.e. the set L of gene locations is regarded as a set (interval) of integers between 1 and $n = \#(L)$, denoted $[1, 2, 3, ..., n]$. A typical crossover γ is given by interchanging alleles, a_l's with b_l's, on a subinterval, $l \in [l_1, l_1 + 1, l_1 + 2, ..., l_2] \subset [1, 2, 3, ..., n]$ for some $1 \leq l_1 \leq l_2 \leq n$ (i.e. $supp(\gamma)$ equals $[l_1, l_1+1, l_1+2, ..., l_2]$ in the so described set L) and/or by composing a few of such transformations. In other words the string may be cut (and recombined) at several (random) locations, where such a cut disengages the corresponding phenotypic features that were linked in the generations before the cut occurred. Sturtevant postulated that the probability of a separating cut is roughly proportional to (or at least monotone increasing with) the distance between the genes and checked out that the available data agree with the idea of linearity.

Thomas Hunt Morgan, who posed the problem to Sturtevant, a 19 year old undergraduate working in his lab, described the result as " one of the most amazing developments in the history of biology"[12]

[11] see http://www.esp.org/foundations/genetics/classical/browse/ for this and other classical papers on genetics.

[12] See
http://www.ias.ac.in/resonance/Nov2003/pdf/Nov2003ArticleInABox.pdf
http://www.esp.org/books/sturt/history/

On the mathematics side, Sturtevant's reasoning may seem to be limited to the banal remark saying that if in a finite metric space the *triangle inequality reduces to equality* on every, properly ordered, triple of points then the metric is *linear*, i.e. inducible from the real line. But this is not exactly what is truly needed as the Sturtevant's linearity is more about the order or, rather the "between" relation, than about metrics.

More interestingly, the idea of Sturtevant suggests the following, novel even from the to-days perspective, way of thinking of geometric structures on a set L that are, according to this point of view, *encoded by probability measures μ on the set 2^L of all subsets $K \subset L$ or by something similar to such measures.*

Typically, one does not have a full direct access to such a measure: the set 2^L is usually too large and individual values $\mu(K), K \in 2^L$ are too small to have any observational meaning. But one may have at one's disposal some quantities – *observed samples* and/or results of controlled specially *arranged experiments* – that provide some information about μ. In the Sturtevant's case, an essential point was designing breeding experiments with the Drosophila flies and below is an example where one relies on an uncontrolled observation.

Reconstruction of the geometry of the physical space from the data provided by the "real world" images. In the model case, the relevant L is the set of pixels on a screen (or of light sensitive cells in the retina of an eye) where we regard L at this stage as just a finite set (of the cardinality from a few thousand to several tens of millions) stripped of any structure, such as the actual geometry of the screen.

An *image* is a partitions of this L into two subsets K_{white} and K_{black}; random observations of the world provide us with a collections of such partitions regarded as samples that are distributed according to some measure μ on 2^L. Reconstructing μ may seem hopeless as we never have enough samples, the set 2^L is huge, but fortunately, the measure (or rather an unknown "something" that we model by a measure) μ governing the real world images is very special: *the closer are the points the more probable they have same color.* In fact, just by "looking" at the images without a preconceived idea of any distance, we can notice that white/black values are strongly positively correlated for *some* pairs (l_1, l_2) in L while for the majority of the pairs there is no correlation at all; then we may interprets this as a manifestation of a distance geometry in L.

Question. Is there, yet unknown, mathematical theory ("multiplication-table type" will do) incorporating these ideas and being useful not only for "specifying parameters" in a (given class of) structures but for also for predicting and/or generating new (classes of) structures?

Mathematics and Pre-mathematics in Biology. Nobody can expect to live through an instance similar to what happened to Hardy and to Weinberg[13], where a purely mathematical thought has clarified a true biology problem: most current applications of mathematics to biology are rather technical and are concerned with a treatment of large amount of dirty data.

However, mathematical thinking may help to generate new useful concepts starting from poorly formalized hints from biology. It may be naive to aim at something comparable to Mendel's formalization of heredity or to Sturtevant's

[13]http://en.wikipedia.org/wiki/Hardy-Weinberg

linearity principle for gene mapping, but having these great examples in mind is encouraging.

Even if not resolving a biology problem one may hope for a flexible formal language for encoding experimental data or/and for something mathematically non-trivial and yet not fully biologically absurd. An example of a simple yet expressive conceptual "book keeping device " is the *evolutionary tree of life* (introduced by Darwin in 1872) while an example of the second kind is *Von Neumann's construction of self-reproducing automata* (somewhere in 1940's), where still there is no theory that would allow a meaningful *formulation preceding* the construction in some "category of models of reproduction", of the creatures inhabiting the living world along with the imaginary mathematical creatures.[14]). Finding such a formulation is an instance of a *pre-mathematical problem*.

References

[1] Craciun, G. Tang, Y., and Feinberg, M., Understanding bistability in complex enzyme-driven reaction networks, Proceedings of the National Academy of Sciences USA, 109, 8697-8702, 2006.

[2] Gorban, A.N. Systems with inheritance: dynamics of distributions with conservation of support, natural selection and finite-dimensional asymptotics, E-print: http://arxiv.org/abs/cond-mat/0405451.

[3] Gorban, A.N., Radulescu,O., Dynamic and static limitation in multiscale reaction networks, revisited, E-print: arXiv:physics/0703278v2 [physics.chem-ph] (2007).

[4] Gromov M., Entropy and Isoperimetry for Linear and non-Linear Group Actions, E-print http://www.ihes.fr/ gromov/topics/grig-may14.pdf.

[5] Liubich Iu. I. Mathematical Structures in Population Genetics, Springer 1992.

[6] Tian J.P., Vojtechovsky, P. Mathematical concepts of evolution algebras in non-Mendelian genetics, E-print http://www.math.du.edu/data/preprints/m0605.pdf

[14] Compare http://en.wikipedia.org/wiki/Self-replication