A more complex class of problems

We extend this in two ways: allow site inhomogeneities, and allow deviation from linkage equilibrium.

Trait: $t^\alpha = \sum_i J_i x_i^\alpha$, with real valued weights J_i.

Selection: acts on a function of this, e.g

- $E = (t - D)^2$; Rattray, 1995,
- $E = V(t)$, with V a two-well function; Shapiro and Prügel-Bennett, 1996, Rogers and Prügel-Bennett, 1999.

We shall use selection weight of $W^\alpha = \exp sE^\alpha$ as before.

This is a much more difficult problem

1. You cannot tell the configuration from the value of the trait.
   - Infer from a Maximum entropy distribution.
2. Selection acts through duplication, so *founder effects* are important.
   - Add an additional macroscopic which measures the similarity of strings in the population; the correlation q.

Correlation:

$$q = \frac{1}{LN^2} \sum_i \sum_\mu \sum_\nu J_i^2 (2x_i^\mu - 1)(2x_i^\nu - 1),$$

$$q = \begin{cases} \frac{1}{L}\sum_i J_i^2 & \sim 1; \quad \text{population completely correlated} \\ \frac{1}{LN}\sum_i J_i^2 & \sim \frac{1}{N}; \quad \text{population completely uncorrelated} \end{cases}$$

Aside on Maximum Entropy Inference

Image a system which can be in a huge number of states: $s_1, s_2, \ldots, s_{N_\Omega} \in \Omega$.

Microscopic knowledge: Probability of finding a member of the ensemble in state s_i, $P(s_i)$. N_Ω numbers.

Macroscopic knowledge: Some average quantity Q.

$$Q = \sum_{s_i \in \Omega} \tilde{Q}(s_i) P(s_i).$$

Problem: Suppose you know Q, and want to know about some other quantity, G,

$$G = \sum_{s_i \in \Omega} \tilde{G}(s_i) P(s_i)$$

if we don't know the P's, how can we do it.

Maximum Entropy Inference: Choose the assignment of the $\{P(s_i)\}$ which is consistent with Q and otherwise has the most "uncertainty" (or assumes the least information).

Uncertainty of a distribution: Given by the entropy

$$H = -\sum_{s_i \in \Omega} P(s_i) \log P(s_i),$$

determines the amount of information required to remove the uncertainty.

Maximum Entropy Distribution: Optimize

$$H = -\sum_{s_i \in \Omega} P(s_i) \log P(s_i) + \lambda \sum_{s_i \in \Omega} \tilde{Q}(s_i) P(s_i)$$

$$\boxed{P(s_i) = \frac{e^{-\lambda \tilde{Q}(s_i)}}{\sum_{s'_i \in \Omega} e^{-\lambda \tilde{Q}(s'_i)}}}$$

The Lagrange multiplier, λ is chosen so that

$$\frac{\sum_{s_i \in \Omega} e^{-\lambda \tilde{Q}(s_i)} \tilde{Q}(s_i)}{\sum_{s'_i \in \Omega} e^{-\lambda \tilde{Q}(s'_i)}} = Q.$$

Maximum Entropy Distribution:

- Occupies almost all of probability distribution space, as $N_\Omega \to \infty$,
- Is an assumption (of microscopic chaos).

Works well if

- Q is “self-averaging; averages of Q are similar to typical ones,
- Q explains most of the non-random aspects of the system (otherwise further macroscopic quantities will be required).

Evolving the correlation due to selection

q relates to $\{\mathbf{x}^\mu\}$ not $\{E^\mu\}$. How to proceed?

1. **Use maximum entropy inference:** In principle, this would be

$$P_{\mathsf{ME}}(\{\mathbf{x}\}|\kappa_1,\ldots) = Z^{-1}e^{-\lambda_1\tilde{\kappa_1}(\{\mathbf{x}\})-\lambda_2\tilde{\kappa_2}(\{\mathbf{x}\})-\ldots}$$

   where Z is the normalization. Lagrange multipliers λ are chosen to give the cumulants we know. This would be used to compute $q^{(s)}$.

   For calculational reasons, we use

$$\boxed{P_{\mathsf{ME}}(\mathbf{x}|\kappa_1,q) = Z^{-1}e^{-\lambda_1\tilde{\kappa_1}(\mathbf{x})-\lambda_2\tilde{q}(\mathbf{x})-\ldots}}$$

   and use to calculate $\kappa_2^{(s)}$.

   This decouples by sites; a distribution with no linkage disequilibrium.

This underestimates correlations due to "founder effects" — selection can only increase fitness due to duplication.

2. **Putting in founder effects:** divide the correlation after selection into two parts:

$$q^{(s)} \quad \rightarrow \quad \underbrace{\tfrac{1}{L}\textstyle\sum_i J_i^2 \displaystyle\sum_{\mu} n^{\mu} n^{\mu}}_{\substack{\text{founder effect} \\ \text{due to duplication}}} \quad + \quad \underbrace{\sum_{\mu}\sum_{\nu} q^{\mu\nu} n^{\mu} n^{\nu}}_{\substack{\text{natural increase} \\ \text{treated via MaxEnt}}}$$

Founder effects: Can be treated as before by expanding around the infinite population,

$$\left\langle \sum_{\mu} \left(\frac{W^{\mu}}{\sum_{\nu} W^{\nu}} \right)^2 \right\rangle \sim \frac{\overline{W^2}}{N\overline{W}^2} + \cdots$$

$$\sim \quad \frac{1}{N}\left(1+\kappa_2 s^2+\ldots\right)$$

Natural increase: Would occur in an infinite population, due to the decrease in state space for higher fitness states. Calculated via MaxEnt.

Crossover Dynamics

Ingredients:

1. Introduce some new crossover variables,

$$C_i^{\mu\nu} = \begin{cases} 1; & \text{if site } i \text{ in } \mu \text{ is replaced by site } i \text{ in } \nu \\ 0; & \text{otherwise} \end{cases}$$

$$\begin{aligned} \mathbf{x}^\mu &\rightarrow \mathbf{x}^\mu + (\mathbf{x}^\nu - \mathbf{x}^\mu) \cdot \mathbf{C}^{\mu\nu} \\ E^\mu &\rightarrow E^\mu + \Delta E^{\mu\nu} \end{aligned}$$

Statistics of C's determined by the crossover model. For simplicity, consider uniform crossover with probability a of crossing at each site.

2. Apply crossover on the *average* mating pool. E.g.

$$\kappa_1^{(c)} = \sum_\mu n^\mu E^\mu + \sum_\mu \sum_\nu n^\mu n^\nu \Delta E^{\mu\nu}$$

Average mating pool: $n^\mu \to \dfrac{W^\mu}{\sum_{\mu'} W^{\mu'}}$

Average over $C^{\mu\nu}$: $\Delta E^{\mu\nu} \to$ $\boxed{\text{work it out}}$

$$\kappa_1^{(c)} = \kappa_1^{(s)}$$

3. Carrying on ...

$$
\begin{aligned}
q^{(c)} &= q \\
\kappa_2^{(c)} &= \kappa_2 - 2a(1-a)\left[\kappa_2 - \frac{1}{4}\left(\sum_i J_i^2 - Lq\right)\right] \\
\kappa_3^{(c)} &= \kappa_3 - 3a(1-a)\left(\kappa_3 - \kappa_3^{\mathrm{LE}}\right) \\
\kappa_4^{(c)} &= \kappa_4 - 2a(1-a)(2-a(1-a))\left(\kappa_4 - \kappa_4^{\mathrm{LE}}\right)
\end{aligned}
$$

The Linkage Equilibrium (LE) cumulants: cumulants of a distribution where every site is drawn independently a binomial:

$$P_{\mathrm{LE}}(x_i^\mu = 1) = \frac{1}{N}\sum_\nu x_i^\nu \equiv \tau_i.$$

But how does the frequency of 1*'s in the current population depend upon the site?* I.e. what is $\tau_i(J_i)$?

We answer that question by use of our maximum entropy distribution:

$$
\begin{aligned}
P_{\mathsf{ME}}(\{x_i^\mu\}) &= \prod_i P_{\mathsf{ME}}(\tau_i) \\
P_{\mathsf{ME}}(\tau_i) &= Z_i^{-1} e^{-\lambda_1 J_i \tau_i - \lambda_2 J_i^2 \tau_i^2}
\end{aligned}
$$

Punch lines about crossover

- Crossover keeps κ_1 and q unchanged,
- Increase in diversity (variance) is related to how similar the strings in the population are (q)
- Higher cumulants are returned towards the LE fixed point which depends upon frequency of 1's at a site as a function of J_i. Computed via MaxEnt.

Mutation Dynamics

Similar to crossover. Introduce mutation variables to the average mating pool, average over them. Mutation rate is γ.

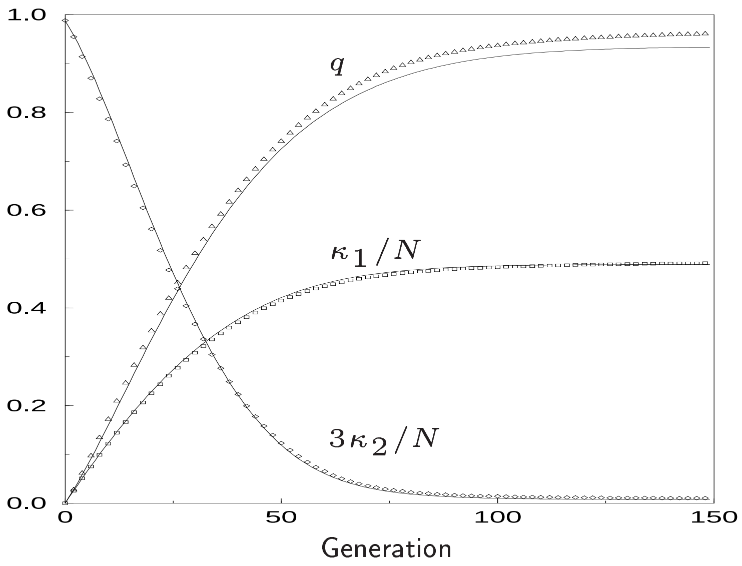
$$
\begin{aligned}
q^{(m)} &= (1-2\gamma)^2 q \\
\kappa_1^{(m)} &= \kappa_1 + 2\gamma(\kappa_1^{\mathrm{random}} - \kappa_1) \\
\kappa_2^{(m)} &= \kappa_2 + 4\gamma(1-\gamma)(\kappa_2^{\mathrm{random}} - \kappa_2)
\end{aligned}
$$

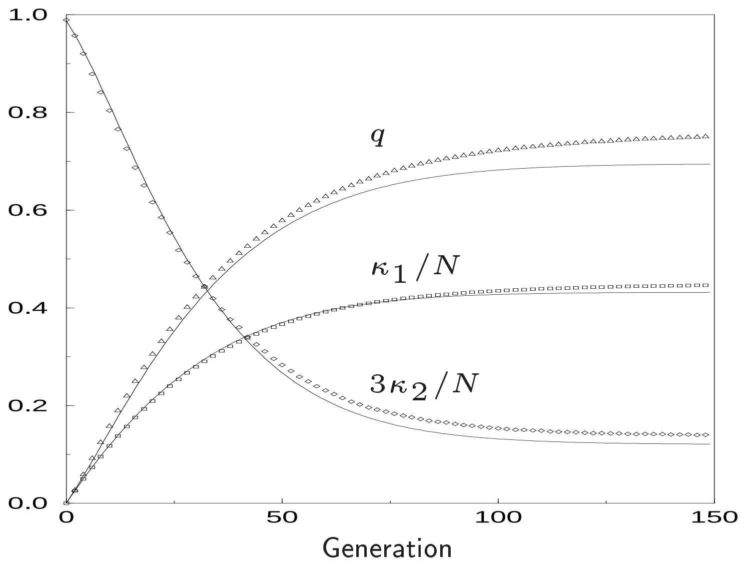
Superscript “random” denotes cumulants of a random population.

The higher cumulants, like those of crossover, relate to the frequency of 1's as a function of J_i, $\tau(J_i)$, which is computed via maximum entropy.

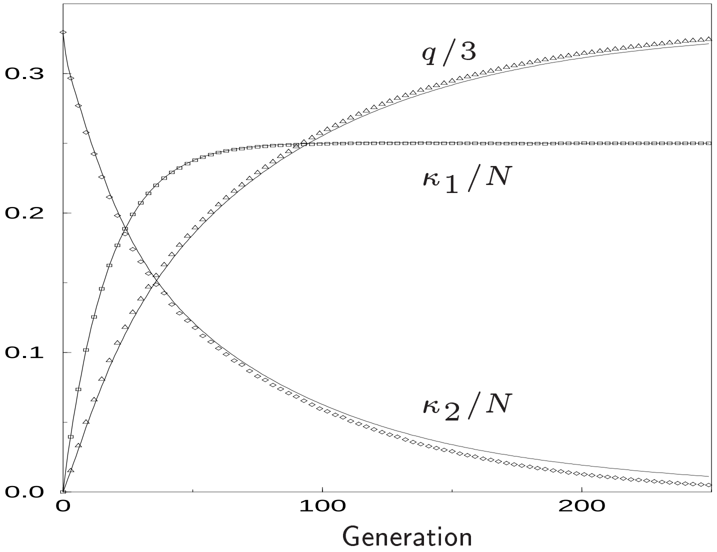
Putting it together

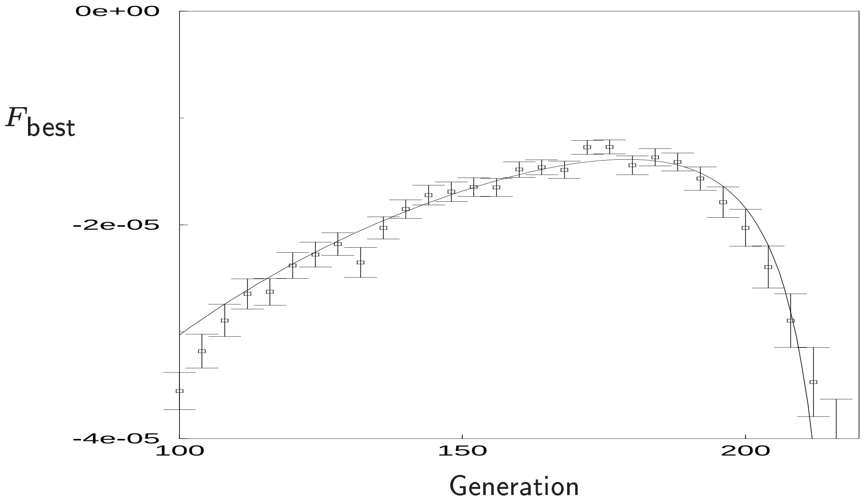
Iterate the process using 4 (or two) cumulants and q.





Results


q/3
κ1/N
κ2/N
0.3
0.2
0.1
0.0
0
100
200
Generation


0e+00
F_best
-2e-05
-4e-05
100
150
200
Generation

The optimum phenotype was $R_{\text{opt}} = N/4$. The other parameters were $N = 80, L = 150, p_m = 0, \beta_s = 0.03$ and uniform crossover was used with $a = 0.5$.

Conclusions

- Evolution of finite populations of long sequences with additive traits can be analyzed by studying the average dynamics of cumulants of the trait distribution.
- In the linkage equilibrium case in the diffusion limit the equations are linear. The dynamics are dominated by a single slow mode. It evolves to the Wright-Fisher fixed point.
- More generally, the cumulants may be near the linkage equilibruim fixed point of crossover. With site inhomogeneities, maximum entropy can be used to determine the linkage equilibrium crossover fixed point. Duplication effects through the correlation can be used to explain deviation from this.

Future work

- Understand how to expand around the linkage equilibrium solution.
- Study how to optimize evolutionary parameters, particularly involving crossover.