Rigorous (and other) results for the NK model

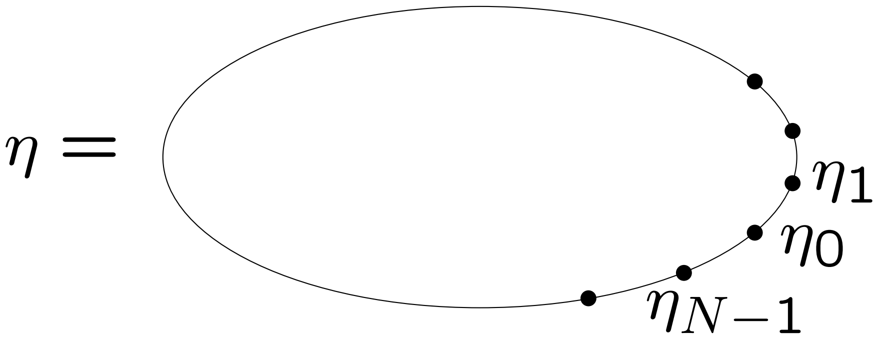
Vlada Limic
University of British Columbia

Plan

- the model
- questions
- biology and applications
- first observations
- general results

General setting

- long sequence of sites (genes in a genome, nucleotides in a DNA, amino acids in a protein)
- each site has a state
- each site contributes to the overall fitness
- fitness contribution depends also on its neighbors
- evolution = process in which the current string changes in order to increase fitness



NK model (Kauffman and Levin (1987))

- sequence with N sites $0, 1, 2, \ldots, N-1$
- site i is in state $\eta_i \in \{0, 1\}$
- i has K neighbors $i+1, i+2, \ldots, i+K$
- modulo N arithmetic
- fitness contribution of site i is

$$\phi_i(\eta_i, \eta_{i+1}, \ldots, \eta_{i+K})$$

- Random variables in the family $\{\phi_i(\nu_i, \nu_{i+1}, \ldots, \nu_{i+K}), i = 0, \ldots, N-1,$ $\nu \in \{0, 1\}^{K+1}\}$ are i.i.d. $U[0, 1]$.

The fitness landscape

The *fitness* of $\eta = (\eta_0, \ldots, \eta_{N-1}) \in \{0,1\}^N$

$$\Phi(\eta) = \sum_{i=0}^{N-1} \phi_i(\eta_i, \eta_{i+1}, \ldots, \eta_{i+K}).$$

Example: $N = 5, K = 1$ then
$\Phi(10001) = \ldots$

generalization: $\phi_i(\eta_i, \eta_{i+1}, \ldots, \eta_{i+K})$ are i.i.d. with distribution F with density f and

$$\int e^{\theta x} f(x)\, dx < \infty \text{ for } \theta \text{ small.}$$

Questions

- How many local maxima are there?
- How many steps does it take to reach a local maximum?
- How different are fitness values at different local maxima?
- How far apart are local maxima?
- What is the value of the global maximum?

In a joint works with Rick Durrett, and with Robin Pemantle, we give qualitative and quantitative answers to some of these questions.

First observations

If $K = 0$ then $\Phi(\eta) = \sum_{i=0}^{N-1} \phi_i(\eta_i)$, so there is one local (and global) maximum.

If $K = N - 1$ then

$$\Phi(\eta) = \sum_{i=0}^{N-1} \phi_i(\eta_i, \eta_{i+1}, \ldots, \eta_{i-1}).$$

For $\eta_1 \neq \eta_2$, $\Phi(\eta_1)$ and $\Phi(\eta_2)$ are independent random variables. So
$P(\eta$ is a local maximum$) =$
$P((0, 0, \ldots, 0)$ is a local maximum$) = \frac{1}{N+1}$,
and there are $2^N/(N + 1)$ local maxima on the average.

For $1 \leq K \leq N - 2$ the fitness landscape is much more complicated.

Biology and other applications

- evolution of species (Sibani-Pedersen (99), Newman (97), Newman-Engelhardt (98),...)
- immune system modelling (Kauffman-Levin (87), Kauffman-Weinberger (89), Macken-Perelson (89), Kauffman(93),...)
- model of organizational change (Levinthal (97))
- evaluation of search algorithms (Merz-Friesleben (99), Park-Nam-Park-Oh (97), Sharpe (00),...)
- model of co-adaptation in international behavior (pol sci?!)

Number of local maxima

Let $G_N = \{(0,0,\ldots,0)$ is a local maximum$\}$.

Let $M_N = \#$ local maxima.

Theorem 1 *[Evans-Steinsalz,Durrett-L]*
For each fixed $K \geq 1$,

$$\lim_{N\to\infty} \frac{\log P(G_N)}{N} = \log \lambda_K,$$

for some $\lambda_K \geq 1/2$.

Since $EM_N = 2^N P(G_N)$,

$$\lim_{N\to\infty} \frac{\log EM_N}{N} = \log 2\lambda_K.$$

Formula for $P(G_N)$

Define $F_2(u) = P(X + Y \le u)$ where X, Y are i.i.d. F.

Suppose $N = 3$, $K = 1$. Event $G_3 = \{(0,0,0)$ is a local maximum$\}$ happens iff

$$\phi_2(0,0) + \phi_0(0,0) \ge \phi_2(0,1) + \phi_0(1,0)$$

$$\phi_0(0,0) + \phi_1(0,0) \ge \phi_0(0,1) + \phi_1(1,0)$$

$$\phi_1(0,0) + \phi_2(0,0) \ge \phi_1(0,1) + \phi_2(1,0).$$

This gives us $P(G_3) =$

$$\int dF(v_0) \int dF(v_1) \int dF(v_2) F_2(v_0 + v_2) F_2(v_1 + v_0) F_2(v_2 + v_1).$$

The formula generalizes!

F general: height of local maxima

Let H_N be the fitness of a randomly (?!) chosen maximum.

Theorem 2 *[Durrett-L] There exist constants μ_H and $\sigma_H > 0$ such that*

$$\frac{H_N}{N} \xrightarrow{p} \mu_H,$$

and

$$\frac{H_N - N\mu_H}{\sqrt{N}} \Rightarrow \mathcal{N}(0, {\sigma_H}^2).$$

"The height of a typical local maximum is $\approx cN$."

F **general**

Let H_N^* be the **fitness of the global maximum**.

Theorem 3 *[Durrett-L] There exist constants μ_H^* and $\sigma_H^* > 0$ such that*

$$\frac{H_N^* - N\mu_H^*}{\sqrt{N}} \Rightarrow \mathcal{N}(0, {\sigma_H^*}^2).$$

Let M_N be the **number of local maxima**.

Theorem 4 *[Durrett-L] There exist constants μ_M and $\sigma_M > 0$ such that*

$$\frac{\log M_N - N\mu_M}{\sqrt{N}} \Rightarrow \mathcal{N}(0, {\sigma_M}^2).$$

cutpoints

For simplicity suppose $K = 1$.

Site i is a *cutpoint* if $\phi_{i-1}(a,1)+\phi_i(1,b) > \phi_{i-1}(a,0) + \phi_i(0,b)$ for all $a, b \in \{0,1\}$.

		1					

- the regions in between the cutpoints are independent!
- H_N^* is obtained by maximizing each of the regions separately and adding up
- M_N is obtained by finding the number of local maxima for each of the regions separately and multiplying

behavior as $N, K \to \infty$

Let $G_{N,K} = \{(0, 0, \ldots, 0)$ is a local max$\}$.

Theorem 5 *[L-Pemantle] If F is standard normal then*

$$\log P(G_{N,K}) = \frac{N}{K}(-\log K + R_{N,K}),$$

where $R_{N,K} \in [-c\sqrt{\log K}, c \log\log K]$ *for some* $c > 0$.

Theorem 6 *[L-Pemantle] If* $N \geq 2(K+1)$ *and F is general then*

$$\begin{aligned}\log P(G_{N,K}) &\leq -(1+o(1))\lfloor \frac{N}{K} - o(1) \rfloor \log K, \\ &\geq -(3+o(1))\lceil \frac{N}{K} \rceil \log K.\end{aligned}$$

We **conjecture** that 3 should be 1.

conclusion

- The NK model is a stochastic model of evolution.
- The fitness landscape is very complicated if $1 \leq K \leq N-2$.
- NK model and modifications are heavily analized and widely used.
- Some of NK model statistical properties have been rigorously studied.