Darwin Channel?

September 14, 2010

Wojciech Szpankowski

Modeling of Darwin Selection

The flow of genetic information can be potentially modeled by two information-theoretic channels, namely the *mutation channel* and the *Darwin channel*. The mutation channel is basically deletion channel, so in this note we try to construct a model for Darwin selection and "survival of the fittest".

Modeling Darwinian selection is a harder problem, since we must deal with a spatio-temporal dynamic process: surviving sequences (genes) re-enter the evolutionary process in time and space. We introduce a (simplified) temporal **Darwin channel** (cf. Figure 1 with feedback that attempts to measure "functional" information transfer in evolution. More precisely, the original input sequence

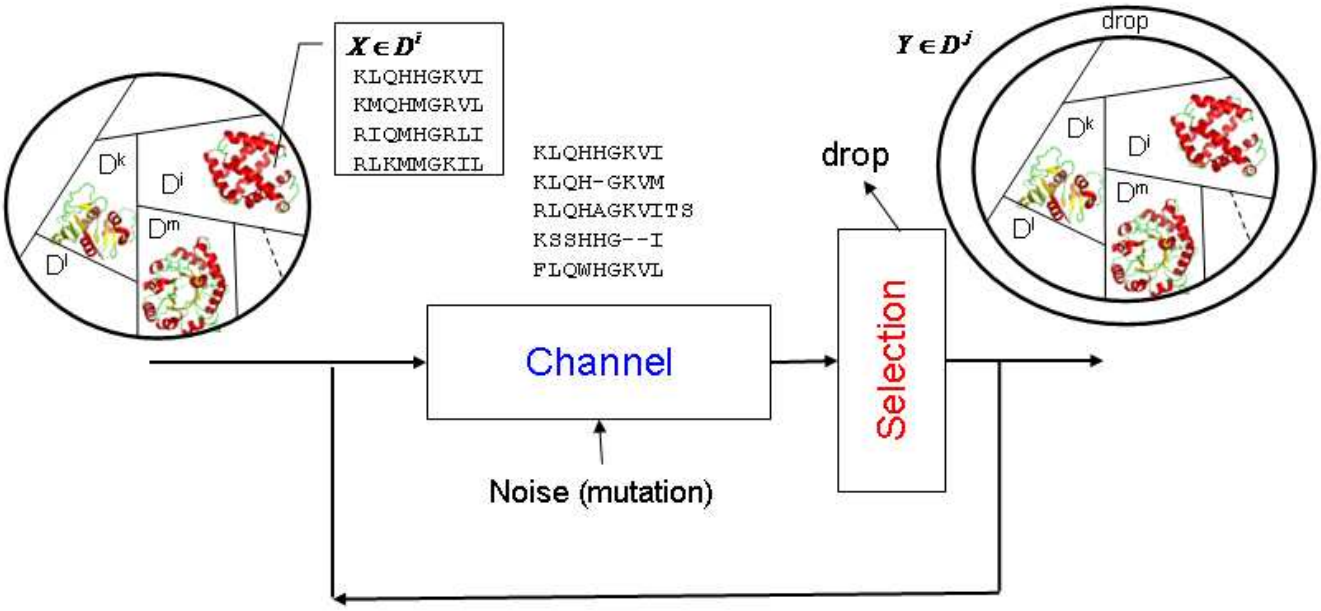

Figure 1: Illustration of the Darwin channels.

X_1^n (e.g., protein or DNA) is restricted to a *constrained* (Darwinian preselected) set $\mathcal{D}_n$ (e.g., globulin protein). This set is partitioned into subsets $\mathcal{D}^i$ of sequences of the same functionality (e.g., by using the scoring function). This partition is represented by a function $F : \mathcal{D}_n \to \{0, 1, \ldots, M-1\}$ such that for all $X_1^n \in \mathcal{D}^I$ we set $F(X_1^n) = I$. In the Darwin channel *mutation* constitutes the *noise*, however, for all practical purposes we assume that substitution is the dominating event. The output sequence Z_1^m may be of random length m, but with substitutions dominating, we shall

assume $m = n$. An output sequence Z_1^n is either assigned to one of the subsets $\mathcal{D}^i$ through F or erased (declare "dead" or not-surviving). In such an information transfer scenario an input sequence is declared to be "functionally surviving" *if both X_1^n and its corresponding output sequence belong to the same functional subset $\mathcal{D}^I$*. Thus the channel preserves functionality. Furthermore, to model temporal behavior, we follow Eigen's observations (i.e., "there are correlations between error rate and genome length") and assume that the error rate is a function of n. To introduce dynamism we assume feedback, hence surviving sequences are sent back as the input to the channel. Note that this model can easily accommodate individual-specific temporal mutations (and associated diseases). Observe that we are in a position to observe the (currently available) output sequences and would like to infer about the past, that is, input sequences.

Is this a realistic model? How to model the "survival of fittest"? What estimates are biologically interesting? *We feel this is not yet the right model, and hope that we come up with a better model during the October workshop.*

Remark 1. Let us assume, we model the flow of information by the mutual information (which may not be the right objective function from biological point of view). In this case, the goal could be to estimate the channel capacity rate C_D, if it exists. It may allow us to estimate the number of functionally interesting sequences on th input side (which is not available to us). If Z_1^n denotes the output sequence, then the capacity C_D is given by

$$C_D = \lim_{n\to\infty} \frac{1}{n} \sup_{P:X_1^n \in \mathcal{D}_n} I(X_1^n, Z_1^n) = \sup_{P:X_1^n \in \mathcal{D}_n} I(X; Z), \tag{1}$$

where $X \in \mathcal{D}_n$ and Z are stationary processes and $I(Z; X)$ represents the mutual information. This is not a trivial mathematical problem (see remark below).

Remark 2. Let us assume that the input sequence X belongs to constrained set $\mathcal{D}$ (surviving protein sequences or their coding genes) and all $\mathcal{D}_i$ are singletons (i.e., contain just one sequence). For simplicity, we consider a binary alphabet with random noise $E_1^n = E_1, \ldots, E_n$ such that $P(E_i = 1) = \varepsilon$ and $P(E_i = 0) = 1 - \varepsilon$ (in real situations ε is often small; e.g., $\varepsilon \approx 10^{-9}$ for mutation in various organisms in gene sequences). Then the problem falls under the paradigm of the *noisy constrained capacity* which we know is a hard problem [1, 2].

References

[1] G. Han and B. Marcus, Asymptotics of the input-constrained binary symmetric channel capacity, *Annals of Applied Probability*, 19, 1063-1091, 2009.

[2] P. Jacquet and W. Szpankowski, Noisy Constrained Capacity for BSC Channels, *IEEE Trans. Information Theory*, 2010.