Subset seed extension to protein BLAST

Anna Gambin Sławomir Lasota Laurent Noé Michał Startek,
Maciej Sykulski and Gregory Kucherov

September 21, 2009

Appendix

Evolution of seed alphabets and seeds

The fitness functions considered in the paper are as follows:

$$\mathrm{fitness}_1(M) = \frac{sens^F(M)}{sens^B(M)}. \tag{1}$$

$$\mathrm{fitness}_2(M) = \begin{cases} 0 & \text{if} \quad sens^B(M) > c \\ sens^F(M) & \text{otherwise} \end{cases}$$

$$\mathrm{fitness}_3(M) = \begin{cases} sens^F(M) & \text{if} \quad sens^F(M) < c \\ \frac{sens^F(M)}{sens^B(M)} & \text{otherwise} \end{cases}$$

$$\mathrm{fitness}_4(M) = \frac{sens^F(M)}{sens^B(M) + c}$$

It turned out that $\mathrm{fitness}_1(M)$ yields unsatisfactory results – the evolution just results in a smallest multiple seed possible, with minuscule foreground and background sensitivity (see Figure 1).

Fitness functions $\mathrm{fitness}_2(M)$, $\mathrm{fitness}_3(M)$ and $\mathrm{fitness}_4(M)$ reflect the trade-off between foreground sensitivity and background sensitivity. The resulting path of evolution for $\mathrm{fitness}_2(M)$ is shown on Figure 2, in this case for $c = 0,002195$. Figure 3 presents the path of evolution for $\mathrm{fitness}_3(M)$ with $c = 0.2$ and Figure 4 for $\mathrm{fitness}_4(M)$ and $c = 0.001$.

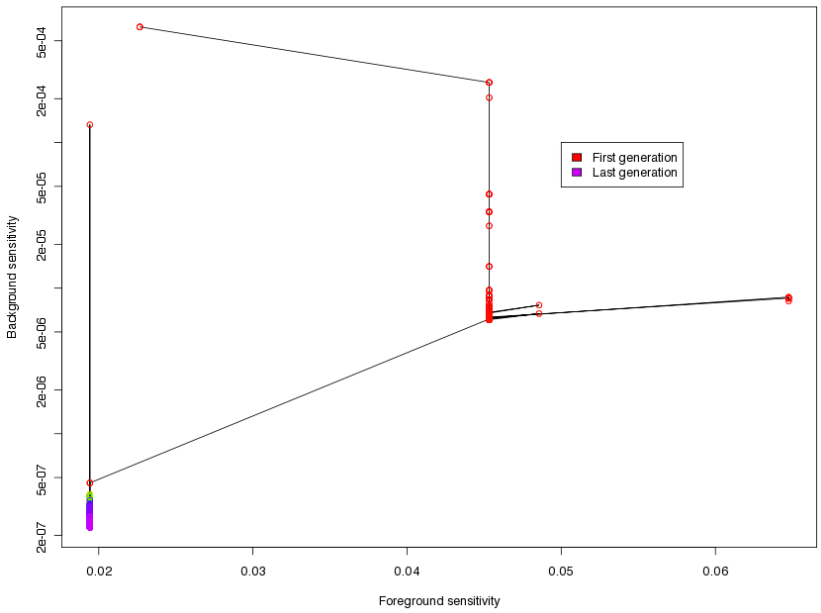


Figure 1: Path of evolution for fitness function $\texttt{fitness}_1$.

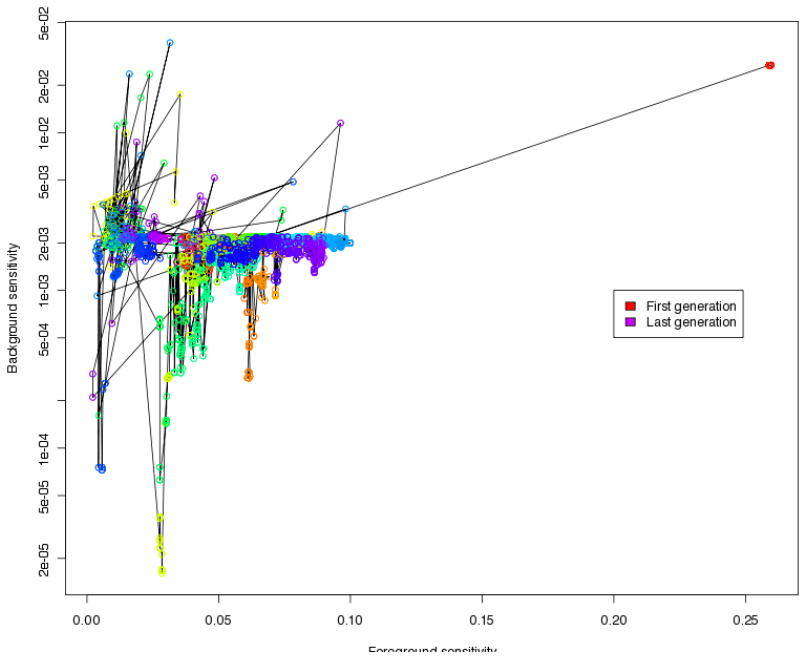


Figure 2: Path of evolution for fitness function $\mathtt{fitness}_2$.

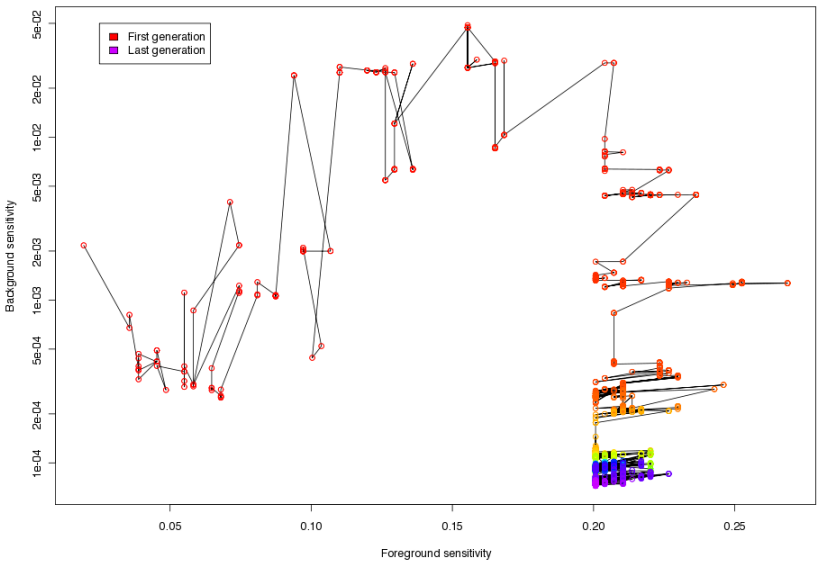


Figure 3: Path of evolution for fitness function $\mathtt{fitness}_3$.

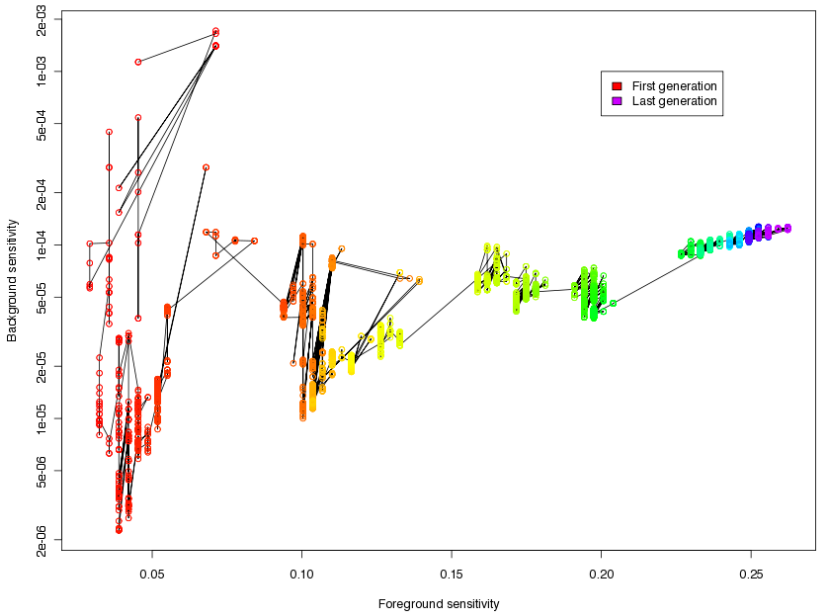
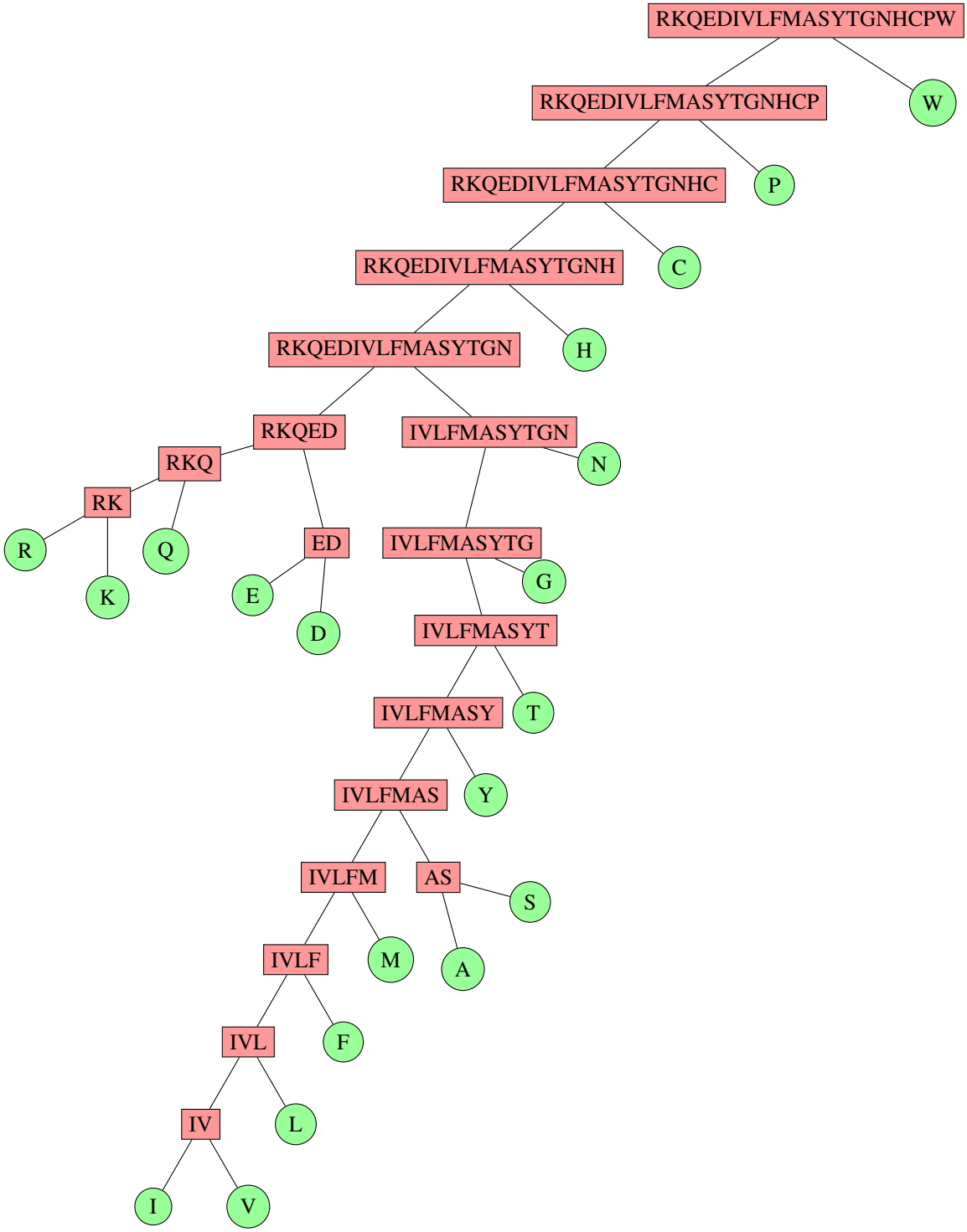


Figure 4: Path of evolution for fitness function $\mathtt{fitness}_4$.

Alphabet of the multiple seed used for experiments

The alphabet contains 8 letters. We provide its representation as a hierarchy of nested equivalence relations over amino acid alphabets. The hierarchical grouping of amino acids is also presented.

$$
\begin{aligned}
\_ &: \{\mathsf{CFYWMLIVGPATSNHQEDRK}\} \\
1 &: \{\mathsf{RKQED}, \mathsf{IVLFM}, A, S, Y, T, G, N, H, C, P, W\} \\
2 &: \{\mathsf{RKQED}, \mathsf{IVL}, F, M, A, S, Y, T, G, N, H, C, P, W\} \\
3 &: \{\mathsf{RKQ}, \mathsf{ED}, \mathsf{IVL}, F, M, A, S, Y, T, G, N, H, C, P, W\} \\
4 &: \{\mathsf{RK}, Q, \mathsf{ED}, \mathsf{IVL}, F, M, A, S, Y, T, G, N, H, C, P, W\} \\
5 &: \{\mathsf{RK}, Q, E, D, \mathsf{IV}, L, F, M, A, S, Y, T, G, N, H, C, P, W\} \\
6 &: \{R, K, Q, E, D, \mathsf{IV}, L, F, M, A, S, Y, T, G, N, H, C, P, W\} \\
\# &: \{R, K, Q, E, D, I, V, L, F, M, A, S, Y, T, G, N, H, C, P, W\}
\end{aligned}
\tag{2}
$$


RKQEDIVLFMASYTGNHCPW
RKQEDIVLFMASYTGNHCP
W
RKQEDIVLFMASYTGNHC
P
RKQEDIVLFMASYTGNH
C
RKQEDIVLFMASYTGN
H
RKQED
IVLFMASYTGN
RKQ
N
RK
R
Q
ED
IVLFMASYTG
K
E
G
D
IVLFMASYT
IVLFMASY
T
IVLFMAS
Y
IVLFM
AS
S
IVLF
M
A
IVL
F
IV
L
I
V