1 Basic definitions

My first attempt at providing a coherent mathematical definition of a repeat element family in the context of a spaced seed.

Definition 1 *A repeat element family descriptor (refd) is a string over the alphabet $\{A, C, G, T, *\}$, describing the contents of any string in the family.*

Definition 2 *Given refds r and r', we will say r' is a slack substring of r (denoted $r' \prec$) if there is a substring r'' of r such that (1) $|r'| = |r''|$, and (2) for all $0 \leq i < |r'|$, either $r'_i = r''_i$, $r'_i = *$, or $r''_i = *$.*

(In otherwords, its a substring, with a potentially different * pattern.) **Note:** I'm not sure if we should be allowing $r''_i = *$.

Definition 3 *We say an refd r matches a Genome G at position i if, for all j such that $r_j \neq *$, $r_j = G_{i+j}$.*

Example: If $r = AA * TT$, and $G = AACTTGGAAGTT$, then r matches G at positions $i = 0$ and $i = 7$. If $G = AAATTT$, then r matches G at positions $i = 0$ and $i = 1$.

Definition 4 *A spaced seed s matches an refd r at position i if, for every $0 \leq j < |s|$, $s_j = 0$ whenever $r_{i+j} = *$.*

Example: If $s = 11011$ and $r = AAAAA * GGG$, then s matches r at positions $i = 0$ and $i = 3$, but not at any other i.

Definition 5 *Given a seed s and refd r, let $M_s(r)$ be the set of values i such that s matches r at i.*

Let $s = 11011$ and $r = AA * CC * GGGGGG$. Then $M_s(r) = [0, 3, 6, 7]$.

Definition 6 *A spaced seed s is consistent with an refd r if for every i, $0 \leq i < |r|$, there is some $i - |s| \leq j \leq i$ such that s matches r at position j.*

In other words: for any position i of the refd, we must be able to match the seed to a position of r such that it then covers position i.

Example: The seed 11011 is compatible with $r = AA * AA * AA$. (The seed matches at positions 0 and 3, and all positions are covered by these two.) But it is not consistent with $AAAAA * * AAAAA$, and there is no seed that can match this string at any position that can cover $i = 5$ or $i = 6$.

Observation 1 *Let L be the sortest sequence of the values in $M_s(r)$. Then s is consistent with r if and only if* $\max_{0<j<|s|-1} L[j+1] - L[j] \leq |s|$.

That is, in the sorted list, every pair of adjacent elements must be within $|s|$ of each other.

Definition 7 *Given a fixed genome G, and fixed spaced seed s, and a fixed value f, we define a elementary repeat family as a set S of genome coordinates, $|S| \geq f$, such that there exists an refd r where:*

- *r matches the sequence of length $|r|$ starting at each element of S. (S is the set of all instances.)*
- *s is consistent with r. (r corresponds to the seed.)*
- *There does not exist an refd r', $r' \prec r$, such that r' is consistent with s and $M_s(r') - M_s(r) \neq \emptyset$. (You cannot have a proper substring of r that describes sequences outside of the instances described by r – minimality.)*
- *There does not exist an refd r', $r \prec r'$, such that s is consistent with r', such that the instances defined by $M_s(r')$ contain all the instances of $M_s(r)$. (Maximality.)*

2 Random Lemmas

The following are things I wanted to try to prove that may or may not be useful.

Lemma 1 *Let s be a seed, r be a refd consistent with s, and r' be an refd consistent with s such that $|r'| < |r|$.*