

Computer generated primer sets

Dr. Peter Taschner

Dr. Hendrik Jan Hoogeboom

Dr. Walter Kusters

Drs. Jeroen Laros

Preprocessing

To reduce space (and time):

$$a \rightarrow 00 = 0$$

$$c \rightarrow 01 = 1$$

$$g \rightarrow 10 = 2$$

$$t \rightarrow 11 = 3$$

Notice that $!a = t$ and $!c = g$.

A string x_0, \dots, x_n is now a number, the next number is x_1, \dots, x_{n+1} .

Linear disk access

By fixating a prefix of the strings, we only search for a part of the possible strings.

AAAA....

AAAC....

...

TTTT....

If we fixate the first digit, we need only $\frac{1}{2}$ of the memory and twice the passes, if we fixate two digits (one nucleotide) we need $\frac{1}{4}$ of the memory and four times the passes.

The output file

First we convert a string to binary form and we reserve space for the score.

ATGCT \Rightarrow

00 11 10 01 11 \Rightarrow

00 00 11 00 10 00 01 00 11 00

After an analysis of length 3, the result could be

00 00 11 00 10 10 01 01 11 11

In DNA terms

A 0 T 0 G 2 C 0 T 3

Which means that ATG is present 3 \times , TGC 1 \times and GCT \geq 4 \times .

Postprocessing

- GC-content
- Bonding energy
- Filtering out simple repeats

GC-content and bonding energy are calculated at the final stage, the simple repeats are filtered out beforehand.

GC-content

ATTAGCAAGAATA has a GC-content of $\frac{3}{13}$.

Determining a large number of these contents goes a lot faster with a sliding window.

string	change	content
ATTA		$\frac{0}{4}$
A TTAG	+1	$\frac{1}{4}$
T TAG C	+1	$\frac{2}{4}$
T AG C A		$\frac{2}{4}$
A G C AA		$\frac{2}{4}$
G CA A G	-1 + 1	$\frac{2}{4}$
C A A GA	-1	$\frac{1}{4}$

And now...

The following is proposed:

- Make the selection process more specific by specifying boundaries.
- Make a selection of 100,000 “acceptable” primers to fill a micro-array.
- Align the products of 50,000 primer pairs against each other.