Problem 1

(a) AAA since $P(O_n = A \mid S_n = T) = P_{AT}$

Prob. of one T

$P(\text{One T} \mid \text{Observe} = AAA)$

$$= \frac{P(O=AAA, S=ATA) + P(O=AAA, S=TAA) + P(O=AAA, S=AAT)}{P(O=AAA)}$$

$P(O=AAA, S=AAA) + P(O=AAA, S=AAT) + P(O=AAA, S=ATA) +$
$P(O=AAA, S=TAA) + P(O=AAA, S=ATT) + P(O=AAA, S=TTA) +$
$P(O=AAA, S=TAT) + P(O=AAA, S=TTT)$

$P(O=AAA, S=ATA) = P(O_1=A \mid S_1=A)\ P(O_2=A \mid S_2=T)\ P(O_3=A \mid S_3=A)\ P(S=ATA)$

Assume every sequence has the same probability in its unobserved state,

$P(S=ATA)$ would be cancelled out

$= P_A \cdot P_{AT} \cdot P_A = P_A^2 P_{AT}$

$\therefore$ $P(\text{one T} \mid \text{observe} = AAA)$

$$= \frac{3 P_A^2 P_{AT}}{P_A^3 + P_A^2 P_{AT} + P_A^2 P_{AT} + P_A^2 P_{AT} + P_A \cdot P_{AT}^2 + P_A \cdot P_{AT}^2 + P_A \cdot P_{AT}^2 + P_{AT}^3}$$

$$= \frac{3 P_A^2 P_{AT}}{P_A^3 + 3 P_A^2 P_{AT} + 3 P_A P_{AT}^2 + P_{AT}^3}$$

(b) $P(O=AAA \mid S=TTT) = \dfrac{P(O=AAA, S=TTT)}{P(O=AAA)}$

$$= \frac{P_{AT}^3}{P_A^3 + 3P_A^2 P_{AT} + 3P_A P_{AT}^2 + P_{AT}^3}$$

(c) $P(O=AGC \mid S=ACC) = \dfrac{P(O=AGC, S=ACC)}{P(O=AGC)}$

$= (P_A \cdot P_{GC} \cdot P_C)$ divided by

$= P_A P_{GC} P_C + P_A P_G P_{GC} + P_{AT} \cdot P_G \cdot P_C$
$+ P_{AT} P_{GC} \cdot P_C + P_A \cdot P_{GC} \cdot P_{CG}$
$+ P_{AT} P_G P_{CG} + P_{AT} P_{GC} P_{CG}$
$+ P_A P_G P_C$

$P(O=AGC) = P(O=AGC, S=ACC) + P(O=AGC, S=AGG) + P(O=AGC, S=TGC) + P(O=AGC, S=TCC) +$
$P(O=AGC, S=ACG) + P(O=AGC, S=TGG) + P(O=AGC, S=TCG) + P(O=AGC, S=AGC)$

(d) AAA exactly one error

would be same as (a)

$$\frac{3 P_A^2 P_{AT}}{P_A^3 + 3 P_A^2 P_{AT} + 3 P_A P_{AT}^2 + P_{AT}^3}$$

(e) AAA. at least one error

$$P(O = AAA \mid \text{at least one error}) = 1 - P(O = AAA \mid \text{no error})$$

$$= 1 - \frac{P(O = AAA,\ S = AAA)}{P(O = AAA)}$$

$$= 1 - \frac{P_A^3}{P_A^3 + 3 P_A^2 P_{AT} + 3 P_A P_{AT}^2 + P_{AT}^3}$$

(f) Observed sequence AAA.

The probability that it is correct = the probability of no error.

$$P(\text{no errors}) = P(S = AAA, E = 000 \mid O = AAA)$$

$$= \frac{P(S_1 = A) \cdot P(S_2 = A) P(S_3 = A) P(E_3 = 0 \mid E_2 = 0) P(E_2 = 0 \mid E_1 = 0) \cdot P(E_1 = 0) \cdot P(O_1 = A) P(O_2 = A) P(O_3 = A)}{P(O = AAA)}$$

$$= \frac{P(S_1 = A) P(S_2 = A) P(S_3 = A) \cdot (1 - P_E)^3 \cdot P_A^3}{P_A^3 + 3 P_A^2 P_{AT} + 3 P_A P_{AT}^2 + P_{AT}^3}$$

(g) exactly 2 errors AAA

$$P(ATT) \quad P(S = ATT, E = 011 \mid O = AAA)$$

$$= \frac{P(S_1 = A) \cdot P(S_2 = A) P(S_3 = A) P(E_3 = 1 \mid E_2 = 1) P(E_2 = 1 \mid E_1 = 0) P(E_1 = 0) P(O = A)^3}{P(O = AAA)}$$

$$= \frac{P(S_1 = A) P(S_2 = A) P(S_3 = A) \, 2P_E \cdot P_E \cdot (1 - P_E) \cdot P_A^3}{P_A^3 + 3 P_A^2 P_{AT} + 3 P_A P_{AT}^2 + P_{AT}^3}$$

$$P(TTA) = \frac{P(S_1 = T) P(S_2 = T) P(S_3 = A) (1 - 2P_E)(2P_E) P_E \cdot P_A^3}{P_A^3 + 3 P_A^2 P_{AT} + 3 P_A P_{AT}^2 + P_{AT}^3}$$

$$P(TAT) = \frac{P(S_1 = T) P(S_2 = A) P(S_3 = T) (P_E)(1 - 2P_E)(P_E) P_A^3}{P_A^3 + 3 P_A^2 P_{AT} + 3 P_A P_{AT}^2 + P_{AT}^3}$$

$$P(\text{exactly 2 errors} \mid O = AAA) = P(ATT) + P(TTA) + P(TAT)$$

(h) $P(E_n = 1) = P(E_n \mid E_{n-1}) \cdot P(E_{n-1} \mid E_{n-2}) \cdot P(E_{n-2} \mid E_{n-3}) \cdots \cdot P(E_1)$

let e denotes the probability of each conditional probability. eg $\sum_0^1 e_1 = P(E_1)$

$\sum_0^1 e_2 = P(E_2 \mid E_1)$ etc.

$\therefore P(E_n=1) = t \cdot \sum_0^1 e_{n-1} \cdot \sum_0^1 e_{n-2} \cdot \sum_0^1 e_{n-3} \cdots \sum_0^1 e_1$

(i) $P(\text{at least 1 error}) = 1 - P(\text{no error})$

$= 1 - (1 - P_E)^n$

Problem 2

(a) since $P(S_i \in (1, \frac{1}{2}\ell_i)) \propto \frac{1}{\ell_i}$

and $P(S_i \in (\frac{1}{2}\ell_i + 1, \ell_i)) \propto \frac{3}{\ell_i}$

The probability of a fragment starting in the first half of the transcript

is $\frac{\ell_i}{2}(c \cdot \frac{1}{\ell_i}) + \frac{\ell_i}{2}(c \cdot \frac{3}{\ell_i}) = 1$

$\therefore P(S_i \in (1, \frac{1}{2}\ell_i)) = \frac{1}{2} \cdot \frac{1}{\ell_i} = \frac{1}{2\ell_i}$ $\quad \Big| \; \frac{\ell_i}{2} \cdot \frac{c}{\ell_i} + \frac{\ell_i}{2} \cdot \frac{3c}{\ell_i} = 1 \Rightarrow \frac{c}{2} + \frac{3c}{2} = 1$

$\frac{4c}{2} = 1 \Rightarrow c = \frac{1}{2}$

(b) The probability of a fragment starting in the second half of the transcript is $P(S_i \in (\frac{1}{2}\ell_i + 1, \ell_i)) = \frac{1}{2} \cdot \frac{3}{\ell_i} = \frac{3}{2\ell_i}$

(c) $E[\tilde{\ell}_i] = \ell_i - E[F] + 1$

$$E[F \mid \ell_i] = \sum_{f=1}^{\ell_i} f \cdot P(F=f)$$

$O=0$

$1 \sim \frac{1}{2}\ell_i \quad \frac{1}{2}\ell_i + 1 \sim \ell_i$

First half $E[F \mid \ell_i] = \sum_{f=1}^{\ell_i - s+1} f \cdot P(F=f)$

$E[\tilde{\ell}_i] = \ell_i - \sum_{f=1}^{\ell_i - s+1} f \cdot P(F=f) + 1$

Second half $E[\tilde{\ell}_i] = \ell_i - \sum_{f=\frac{1}{2}\ell_i+1}^{\ell_i - s+1} f \cdot P(F=f) + 1$

$O=1$

$\frac{1}{2}\ell_i + 1 \sim \ell_i \quad 1 \sim \frac{1}{2}\ell_i$

first half $E[\tilde{\ell}_i] = \ell_i - \sum_{f=\frac{1}{2}\ell_i+1}^{\ell_i - s+1} f \cdot P(F=f) + 1$

second half $E[\tilde{\ell}_i] = \ell_i - \sum_{f=1}^{\ell_i - s+1} f \cdot P(F=f) + 1$

$\therefore$ The effective length of any distribution F would be

$$E[\tilde{\ell}_i] = \ell_i - \sum_{1}^{\ell_i} \sum_{0}^{1} \sum_{1}^{\ell_i - s+1} f \cdot P(F=f \mid S=s, O=o) \cdot P(S) \cdot P(O) + 1$$

```python
#E[F|l] = sum of f = 1 to li of f * P(F = f)
import math
def effective_length_(transcript_length, miu, sigma):
    expected_F_on_l = 0
    sum = 0
    for f in range(1, transcript_length + 1):
        sum += (math.exp((f - miu)**2 /(2 * sigma**2)* (-1)))

    normalizing_const = 1.0 / sum

    for f in range(1, transcript_length + 1):
        expected_F_on_l += f * normalizing_const * (math.exp((f - miu)**2 /(2 * sigma**2)* (-1)))

    expected_effective_length = transcript_length - expected_F_on_l + 1

    return expected_effective_length
```

Python

```python
#(a) li = 1000, μ = 200, σ = 20,

print(effective_length_(1000, 200, 20))
```

Python

```
800.9999999999999
```

```python
#(b) li = 1000, μ = 200, σ = 100.

print(effective_length_(1000, 200, 100))
```

Python

```
795.418280719455
```

(e) The values decreased a little. It's a reasonable behavior b/c as value of σ gets bigger, the value of the probability mass function will get smaller.