Why?

For $k > 2$ restriction R can not be satisfied as partial nodes are always Q-nodes whose front may only be inverted. So the resulting permutation would contain empty (do not belong to R) in between full nodes (belong to R). This contradicts our definition of a restriction forcing all elements in R to be neighbored.

In all other possible cases for P-nodes full and empty nodes would be nested making satisfaction impossible for the same reasons as above.

Q-nodes:
Rules Q0 to Q3, with the restriction on Q3 that x has to be the root of the smallest subtree of T containing all of R.
For the same reasons as before we only have to consider Q-nodes with at most two partial children.

Consecutive ones problem: Create $\mathcal{R} := \{R_1, \ldots, R_n\}$, where R_i contains exactly those column numbers where the STS matrix has entry 1 in row i.
Satisfying $R_i \Rightarrow$ all ones in row i are consecutive.
Starting with the universal PQ tree for $\{1, 2, \ldots, m\}$ (column numbers) and reducing to satisfy restrictions R_i, $i = 1, 2, \ldots, m$, one after another we get the empty tree, if the matrix does **not** have the consecutive ones property. Else we get a PQ tree T whose permutation set $Perm(T)$ represents exactly those permutations that transform the given matrix into consecutive ones form.

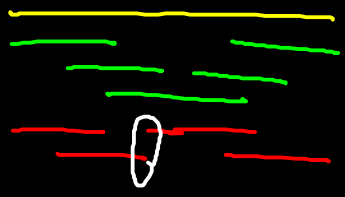
Example: $\mathcal{R} = \{\{A, B\}, \{A, B, C, D\}, \{A, D\}, \{D\}\}$.

Theorem

Let A be a $n \times m$-matrix over $\{0, 1\}$ and let k be the number of ones in A. Then the previously mentioned procedure solves the consecutive ones problem in time $\mathcal{O}(n + m + k)$. □

It is important to implement the rules efficiently.

Stochastic modeling: How good do randomly chosen fragments cover a molecule?



Definition

Let $A : \mathbb{R}_0 \to \mathbb{N}$ a nondecreasing function satisfying $A(0) = 0$, where $A(t)$ describes the number of events until time t. Then we have a poisson process with rate λ, if

(1) $\Pr[A(s+t) - A(s) = n] = \exp(-\lambda \cdot t)\frac{(\lambda \cdot t)^n}{n!}$ *and*

(2) *the distribution of the number of events is stationary, i.e. it depends only on the length but not on the position of a given interval.*

Theorem

Let A be a poisson process.

a) *The expected number $\mathbb{E}[A(t)]$ of events in an interval of length t satisfies $\mathbb{E}[A(t)] = \lambda \cdot t$.*

b) *Let T_n be the time between the $(n-1)$-th and the n-th event. Then*

$$\Pr[T_1 > t] = \Pr[T_n > t] = \exp(-\lambda \cdot t).$$

Beweis: $\mathbb{E}[A(t)] = \sum_{k \geq 0} k \cdot \Pr[A(t) = k]$

$= \sum_{k \geq 1} k \cdot \exp(-\lambda t) \cdot \frac{(\lambda t)^k}{k!}$

$= \exp(-\lambda t) \sum_{k \geq 1} k \frac{(\lambda t)^k}{k!} = \exp(-\lambda t) \cdot \sum_{k \geq 1} \frac{(\lambda t)^k}{(k-1)!}$

$= \exp(-\lambda t)\,(\lambda t) \sum_{k \geq 1} \frac{(\lambda t)^{k-1}}{(k-1)!}$ $\qquad e^x = \sum_{k \geq 0} \frac{x^k}{k!}$

k, $k-1$, 0

$= \exp(-\lambda t)\,(\lambda t) \cdot \exp(\lambda t) = \lambda t$

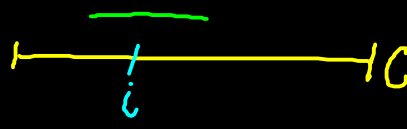
b) $P_1[T_1 > t] = P_1[T_n > t]$ da Prozess stationär.

und $P_1[T_1 > t] = P_1[A(t) = 0]$

Setzen wir in (1) $n = s = 0$ so folgt

$$P_1[T_n > t] = P_1[T_1 > t] = P_1[A(t) = 0]$$
$$= \exp(-\lambda t).$$

Model: Assuming fragments of length L are cut from multiple copies of a DNA molecule of length C randomly and independently. Then for i any position of the molecule



$$\Pr[i \text{ is covered by randomly chosen fragment}] = \frac{L}{C}$$

holds and thus

$$\left(1 - \frac{L}{C}\right)^N \approx \exp\left(-\frac{L \cdot N}{C}\right) \qquad (1)$$

is the probability that i is not covered by any of the N fragments.

Poisson process?

We let $\lambda := \frac{L}{C}$ and ask for the probability of exactly n of the N fragments covering position i. This probability is given by

$$\begin{aligned}\binom{N}{n} \cdot \left(\frac{L}{C}\right)^n \cdot \left(1 - \frac{L}{C}\right)^{N-n} &\approx \frac{N!}{n! \cdot (N-n)!} \lambda^n \cdot \exp(-\lambda \cdot N) \\ &= \frac{N!}{N^n \cdot (N-n)!} \exp(-\lambda \cdot N) \cdot \frac{(\lambda \cdot N)^n}{n!} \\ &\approx \exp(-\lambda \cdot N) \cdot \frac{(\lambda \cdot N)^n}{n!}.\end{aligned}$$

So the number of fragments covering a fixed position is approximately poisson distributed with rate $\lambda = \frac{L}{C}$. (This statement holds for $n \ll N \ll C$ and $L \ll C$.)

The expected number of fragments covering position i is thus $R := \lambda \cdot N = \frac{L \cdot N}{C}$. R is called *redundancy* of the fragment set.

Corollary

The expected number of positions not covered is (approximately) given by

$$\exp\left(-\frac{L \cdot N}{C}\right) \cdot C = \exp(-R) \cdot C.$$

□

Definition

Let $\mathcal{F}$ be a set of fragments of length L and $\Theta \in [0, 1]$. We take $\mathcal{F}$ as vertices of a graph and connect two fragments $f_1, f_2 \in \mathcal{F}$ by an indirected edge, if a suffix (prefix) of f_1 of length at least $\Theta \cdot L$ is a prefix (suffix) of f_2 (overlap). We get an undirected graph whose connected components are called Θ-islands.

Intuition: Fragments with only small overlap should not be considered overlapping.

How many Θ-islands are to be expected?

Lemma

Let $\Theta \in [0,1]$ and redundancy R be given and let N be the number of randomly chosen fragments. Then

$$N \cdot \exp(-R \cdot (1-\Theta))$$

is (approximately) the expected number of Θ-islands.

Es sei i eine feste Position des Moleküls.

Wir definieren

$J(x) = Pr[$ Positionen i und $i+x \cdot L$ gehören zu Keinem gemeinsamen Fragment$]$

(zufällig)

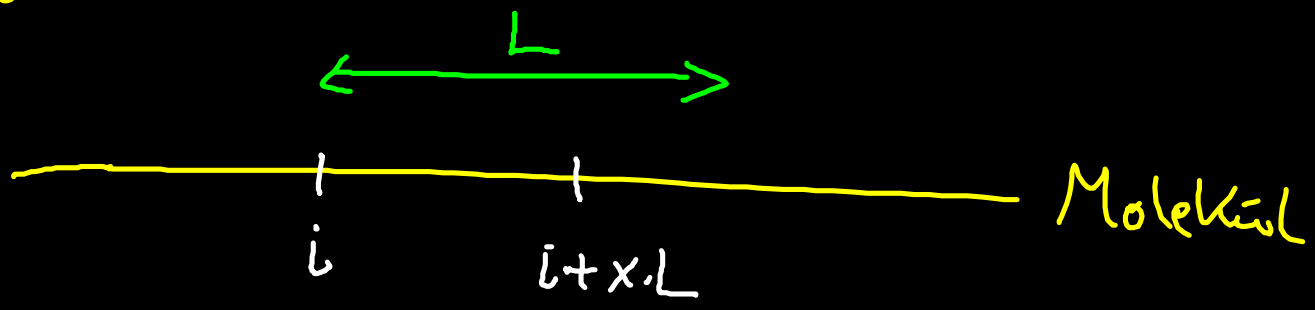
Behauptung

$$J(x) \approx \begin{cases} \exp(-R(1-x)) & \text{für } 0 \le x \le 1; \\ \emptyset & \text{sonst} \end{cases}$$

mit $R = \lambda \cdot N = \frac{L \cdot N}{C}$ die Redundanz

Betrachte zufällig gewähltes Fragment der

Länge L



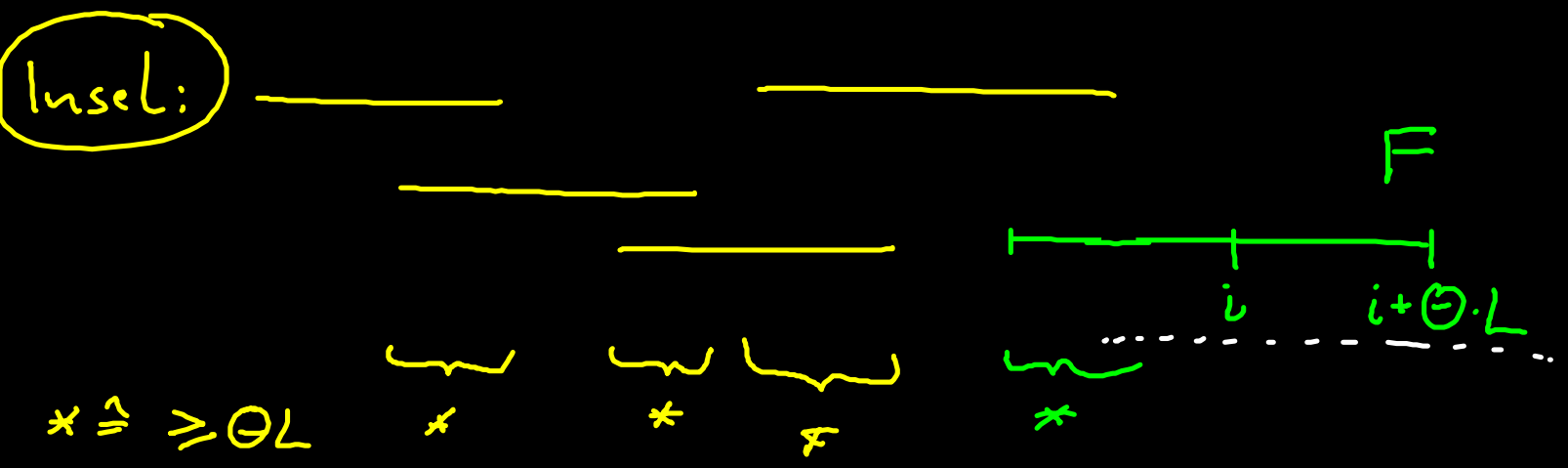
Anzahl günstiger Möglichkeiten für die Lage des zufällig gew. Fragments, so dass i und $i+x.L$ überdeckt werden ist $(1-x).L$

$\Rightarrow Pr[i \text{ und } i+x.L \text{ werden von zufällig gew. Frag. überdeckt}]$

$$= \frac{(1-x).L}{C}$$

Setze $\lambda = \frac{(1-x)L}{C}$, dann wie zuvor.

Insel:



Beobachtung: F ist das einzige Fragment, das

sowohl i als auch i+ΘL überdeckt

$\mathbb{E}$ # Θ-Inseln $\hat{=}$ $\mathbb{E}$ # am weitesten rechts beh. Fragmente

$= \mathbb{E}$ # Fragmente ohne Par. i und i+ΘL überdecken

$= N \cdot J(\Theta) \approx N \cdot \exp(-R(1-\Theta))$ □

Example: We consider the case of a molecule with 10^8 bases to be mapped. We assume that a library of 10000 fragments has been created, each around 50000 bases long. In this case $R = \frac{5 \cdot 10^4 \cdot 10^4}{10^8} = 5$ and for Θ small enough $N \cdot \exp(-R \cdot (1-\Theta)) \approx 10^4 \cdot \exp(-5) = 67.37946999\ldots$ many islands are to be expected.

Shotgun sequencing and fragment assembly:

Definition

Let D be a DNA molecule to be sequenced and $S = \{s_1, \ldots, s_n\}$ the set of words (fragment sequences), observed at a shotgun sequencing of D. Then the fragment assembly problem *is to determine (algorithmically) the arrangement of the words from S corresponding to their original positions in D.*