Optical maps

Introduction

Optical mapping is a technology that allows us to experimentally determine the relative position of certain landmarks (usually restriction sites - places where a specific restriction enzyme can cut the DNA - usually recognized by a specific 6-8 bp sequence) along a stretch of DNA. Unlike sequencing, optical maps provide long range information - the fragments being mapped are usually on the order of several 100s of kbp - however the information provided is sparse. Thus optical mapping is useful as a complement to sequencing. Also, it can be used as a cheaper way (than sequencing) of getting information about the global structure of a genome.

Computationally an optical map is just an ordered list of sizes, together with estimates of the error in the size estimates, representing the list of gaps between adjacent restriction sites.

Mapping algorithm

I'll focus here on just the problem of aligning an experimentally determined optical map to an *in silico* optical map constructed, e.g. from the output of an assembler.

Formally, the experimental map is represented as the array:

$$emap = \{(o_k, s_k), k = 1, n\}$$

where o_k and s_k are the size and standard deviation for fragment k.

The *in silico* map is:

$$ismap = \{e_k, k = 1, m\}$$

where e_k is the size of the corresponding fragment

Aligning the two maps can be performed pretty easily using a dynamic programming algorithm similar to the sequence alignment algorithm.

Specifically, V[i,j] is the score of aligning the first i fragments from the experimental map to the first j fragments from the *in silico* map.

The recurrence equation is:

$V[i,j] = \min_{k<i,l<j} \{ V[k,l] + score(k..i, l..j)\}$

where score(k..i, l..j) is a score of how well the set of fragments between k..i, and l..j, match each other. This score can be defined as a combination of a X^2 score and a penalty for missed sites:

$$score(k..i, l..j) = \frac{\left(\sum_{s=k}^{i} o_s - \sum_{t=l}^{j} e_t\right)^2}{\sum_{s=k}^{i} s_s^2} + C(i - k + l - j)$$

where C is a constant that can be used to tune the contribution of the two components.

Interesting research directions

- Can optical maps be used to guide genome assembly?
- How do you efficiently align maps to an already sequenced genome to identify structural variants?
- How do you efficiently align two maps to each other to identify structural differences?