Ingenieurspraxis

Coding for DNA Microarrays

Gene expression analysis


Measuring gene expression is an important part of many life sciences, as the ability to quantify the level at which a particular gene is expressed within a cell, tissue or organism can provide a lot of valuable information. For example, measuring gene expression can:

- Identify viral infection of a cell (viral protein expression).
- Determine an individual's susceptibility to cancer (oncogene expression).
- Find if a bacterium is resistant to penicillin (beta-lactamase expression).

Similarly, the analysis of the location of protein expression is a powerful tool, and this can be done on an organismal or cellular scale. Investigation of localization is particularly important for the study of development in multicellular organisms and as an indicator of protein function in single cells. Ideally, measurement of expression is done by detecting the final gene product (for many genes, this is the protein); however, it is often easier to detect one of the precursors, typically mRNA and to infer gene-expression levels from these measurements.

An approach for gene expression analysis is the [hybridization microarray](https://en.wikipedia.org/wiki/Hybridization_microarray). A single array or "chip" may contain probes to determine transcript levels for every known gene in the genome of one or more organisms.

Here is a video explaining the concepts of gene expression analysis and DNA microarrays: [https://www.youtube.com/watch?v=Hv5f1U0sE0s](https://www.youtube.com/watch?v=Hv5f1U0sE0s)

Relation to coding

The generation of the DNA microarray bears several interesting challenges from a coding perspective. The microarray needs to be generated, which means that artificial DNA strands are synthesized. Many of these DNA strands are synthesised simultaneously. However, only one of the four bases of the DNA (A, C, G, T) can be written at once. The challenge is to select the best possible sequence in which the bases are written to minimize the overall running time. This problem is called finding the Shortest Common Supersequence (SCS) and is unfortunately np-hard ([https://en.wikipedia.org/wiki/Shortest_common_supersequence](https://en.wikipedia.org/wiki/Shortest_common_supersequence)). Furthermore, there are more than one possible probes that can be used to signal a specific gene. Hence, this is another parameter to optimize.

Problem description

This thesis aims to analyze the problem of finding the best probes for a DNA microarray and implement an algorithm to run small-scale experiments.

This should only roughly introduce the problem. No need to understand everything or review all references. If you are interested, please reach out to me, and we can discuss a suitable direction for you.

Related references


**Prerequisites**

Good knowledge and high interest in mathematics, especially

- Linear algebra
- Combinatorics

Good programming knowledge is required for this topic.

**Contact**

frederik.walter@tum.de

**Advisors**

Frederik Walter