Design and Optimization ofUniversal DNA Arrays

Ion Mandoiu

CSE Department & BME Program
University of Connecticut
DNA Microarrays

• Exploit Watson-Crick complementarity to simultaneously perform a large number of **substring tests**

• Used in a variety of high-throughput genomic analyses
  – Transcription (gene expression) analysis
  – Single Nucleotide Polymorphism (SNP) genotyping
  – Genomic-based microorganism identification
  – Alternative splicing, ChIP-on-chip, tiling arrays,…

• Common microarray formats involve **direct hybridization** between labeled DNA/RNA sample and DNA probes attached to a glass slide
Universal DNA Arrays

• Limitations of direct hybridization formats:
  – Arrays of cDNAs: inexpensive, but can only be used for transcription analysis
  – Oligonucleotide arrays: flexible, but expensive unless produced in large quantities

• Universal DNA arrays: “programable” arrays
  – Array consists of application independent oligonucleotides
  – Detection carried by a sequence of reactions involving application specific primers
  – Flexible AND cost effective

• Universal array architectures: DNA tag arrays, APEX arrays, SBE/SBH arrays
Mix tag+primer probes with genomic DNA

Solution phase hybridization

Solid phase hybridization

Single-Base Extension
Tag Hybridization Constraints

(H1) Tags hybridize strongly to complementary antitags
(H2) No tag hybridizes to a non-complementary antitag
(H3) Tags do not cross-hybridize to each other

Tag Set Design Problem: Find a maximum cardinality set of tags satisfying (H1)-(H3)
Hybridization Models

• Hamming distance model, e.g., [Marathe et al. 01]
  – Models rigid DNA strands

• LCS/edit distance model, e.g., [Torney et al. 03]
  – Models infinitely elastic DNA strands

• c-token model [Ben-Dor et al. 00]:
  – Duplex formation requires formation of nucleation complex between perfectly complementary substrings
  – Nucleation complex must have weight ≥ c, where \( \text{wt}(A) = \text{wt}(T) = 1, \text{wt}(C) = \text{wt}(G) = 2 \) (2-4 rule)
c-h Code Problem

• **c-token**: left-minimal DNA string of weight $\geq c$, i.e.,
  - $w(x) \geq c$
  - $w(x') < c$ for every proper suffix $x'$ of $x$

• A set of tags is a **c-h code** if
  (C1) Every tag has weight $\geq h$
  (C2) Every c-token is used at most once

---

**c-h Code Problem** [Ben-Dor et al.00]

*Given $c$ and $h$, find maximum cardinality c-h code*

[Ben-Dor et al.00] give approximation algorithm based on DeBruijn sequences
Periodic Tags [MT05]

- Key observation: c-token uniqueness constraint in c-h code formulation is too strong
  - A c-token should not appear in two different tags, but **can be repeated in a tag**
  - Periodic tags use fewer c-tokens!

→ Tag set design can be cast as a cycle packing problem
c-token factor graph, c=4 (incomplete)
Cycle Packing Algorithm

1. Construct c-token factor graph $G$
2. $T \leftarrow \{\}$
3. For all cycles $C$ defining periodic tags, in increasing order of cycle length,
   - Add to $T$ the tag defined by $C$
   - Remove $C$ from $G$
4. Perform an alphabetic tree search and add to $T$ tags consisting of unused c-tokens
5. Return $T$

– Gives an increase of over 40% in the number of tags compared to previous methods
## Experimental Results

<table>
<thead>
<tr>
<th>$h$</th>
<th>$c$</th>
<th>One $c$-token copy</th>
<th>Multiple $c$-token copies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LP approx</td>
<td>Tree search</td>
</tr>
<tr>
<td></td>
<td></td>
<td>tags</td>
<td>$c$-tokens</td>
</tr>
<tr>
<td>$h \geq 28$</td>
<td>4</td>
<td>3</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>8</td>
<td>151</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>22</td>
<td>391</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>63</td>
<td>1083</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>182</td>
<td>2996</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>515</td>
<td>8025</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>1491</td>
<td>22183</td>
</tr>
</tbody>
</table>
More Hybridization Constraints...

- Enforced during **tag assignment** by
  - Leaving some tags unassigned and distributing primers across multiple arrays [Ben-Dor et al. 03]
  - Exploiting availability of multiple primer candidates [MPT05]
## Herpes B Gene Expression Assay

### GenFlex Tags

<table>
<thead>
<tr>
<th>T_m</th>
<th># pools</th>
<th>Pool size</th>
<th>500 tags</th>
<th>1000 tags</th>
<th>2000 tags</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td># arrays</td>
<td>% Util.</td>
<td># arrays</td>
</tr>
<tr>
<td>60</td>
<td>1446</td>
<td></td>
<td>1</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>67</td>
<td>1560</td>
<td></td>
<td>1</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>70</td>
<td>1522</td>
<td></td>
<td>1</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

### Periodic Tags

<table>
<thead>
<tr>
<th>T_m</th>
<th># pools</th>
<th>Pool size</th>
<th>500 tags</th>
<th>1000 tags</th>
<th>2000 tags</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td># arrays</td>
<td>% Util.</td>
<td># arrays</td>
</tr>
<tr>
<td>60</td>
<td>1446</td>
<td></td>
<td>1</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>67</td>
<td>1560</td>
<td></td>
<td>1</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>70</td>
<td>1522</td>
<td></td>
<td>1</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>
New SBE/SBH Assay

Primer

<table>
<thead>
<tr>
<th>AA</th>
<th>AC</th>
<th>CC</th>
<th>CA</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT</td>
<td>AG</td>
<td>CG</td>
<td>CT</td>
</tr>
<tr>
<td>TT</td>
<td>TG</td>
<td>GG</td>
<td>GT</td>
</tr>
<tr>
<td>TA</td>
<td>TC</td>
<td>GC</td>
<td>GA</td>
</tr>
</tbody>
</table>
SBE/SBH Throughput \((c=13, r=5)\)

See poster for more details
Conclusions and Ongoing Work

- Combinatorial algorithms yield significant increases in multiplexing rates of universal DNA arrays
  - New SBE/SBH architecture particularly promising based on preliminary simulation results
- Ongoing work:
  - Extend methods to more accurate hybridization models, e.g., use NN melting temperature models
  - More complex (e.g., temperature dependent) DNA tag set non-interaction requirements for DNA self/mediated assembly
  - Probabilistic decoding in presence of hybridization errors
Acknowledgments

• UCONN Research Foundation
• Claudia Prajescu
• Dragos Trinca