CSE91
Early detection of cancers using computation

Vineet Bafna, UCSD
A combinatorial puzzle

- Given: a graph with nodes laid out on a line
- Select a subset of nodes
  - 2 nodes that are connected by an edge cannot both be in the subset
  - No adjacent pair is apart by more than 1

```
1   2   3   4   5   6   7   8   9   10  11
```

11/5/09

Bafna
How hard is the puzzle

• Number of nodes in the graph: \( n \) (5000)
• Total length of the line: \( k \) (500)
• Minimum number of nodes:

\[
\binom{n}{k} = \binom{5000}{500} = 10^{500}
\]

• Number of ways of selecting subsets
Today

• We talk about smarter ways to solve this puzzle
• And,
• we show how to use the solution for early detection of cancer.
Gleevec?
The Gleevec story

• The Food and Drug Administration approved a new kind of anti-cancer drug yesterday after a swift, two-and-a-half-month review of clinical data. Drugs accepted for priority review usually take six months to approve.

• Gleevec's quick approval rests on its immediate and striking results in treating patients with chronic myelogenous leukemia, one of the four main types of leukemia. In one trial of 54 patients, Gleevec caused their blood counts to revert to normal in more than 90 percent of cases and cancerous cells to disappear altogether in 13 percent, with generally few side effects.

• ...Gleevec was the "first molecularly targeted drug" and showed "the incredible value of developing a drug against a target" when the errant component of a cancer cell could be identified.
Cancer?

- In normal cells, when the DNA is damaged beyond repair, the cells commit "suicide" (apoptosis).
- In cancer, these cells proliferate, invade other tissues.
The Science behind Gleevec

Fusions
- observed in leukemia, lymphoma, and sarcomas
  - “Philadelphia Translocation”
  - Drugs target this fusion protein

1) Increased proliferation
2) Decreased apoptosis
3) Disturbed interaction with the cell’s extracellular matrix
What did we learn?

• Something in the DNA of a tumor cell changes.
  – This something is a large ‘rearrangement’ of DNA, like deletion, translocation,…

• The rearrangement changes genes responsible for growth, apoptosis etc., eventually causing cancer.
Fluorescent in situ hybridization

- Cancer genomes show extensive structural variation
DNA

• Think of a DNA as a string over A,C,G,T
• Actually, two complementary strings that stick to each other (A<->T, C<->G)

ACGGATACGATCAGATAAGTAGATCGTGTGTGGG
TGCCTATGCTAGTCTATTCATCTAGCACACACACCC
DNA deletion

- In a cancer, some region might get deleted.
- If we can detect the deletion, we can detect the cancer.

ACGGATACGATCGATCAGATCGTGTGTGGG
TGCCTATGCTAGTCTATTCATCTAGCACACACACCC

ACGGATACGATCAGATCGTGTGTGGG
TGCCTATGCTAATCTAGCACACACCC

ACGGATACGATCAGATTAGATCGTGTGTGGG
TGCCTATGCTAATCTAGCACACACACCC

11/5/09  Bafna
Assaying for Rare Variants

• How about the following?
• Take a random piece of DNA
  – Sequence it.
  – Compare to a reference.
  – Check if there is a deletion.
• Any problems?
Assaying for Rare Variants

- Most samples are not from the deletion.
- We can keep sequencing and never detect the deletion.
- Also, in the early stages of cancer, most DNA is normal.
Polymerase Chain Reaction

- PCR is a technique for amplifying and detecting a specific portion of the genome
- Amplification takes place if the primers are ‘appropriate’ distance apart (<2kb)
PCR is DNA-copier

ACGGATACGATCAGATAAGTAGATCGTGTGTGGG
TGCCTATGCTAGTCTATTCATCTAGCACACACACCC

ACGGATACGATCAGATAAGTAGATCGTGTGTGGG
TGCCTATGCTAGTCTATTCATCTAGCACACACCC

heat
TGCCTATGCTAGTCTATTCATCTAGCACACACCC

primer
TGCCTA'TGCCTATGCTAGTCTATTCATCTAGCACACACCC
ACGGATACGATCAGATAAGTAGATCGTGTGTGGG

2000 bp can be amplified
Assaying for Rare Variants

• PCR can be used to assay for a given genomic abnormality, even in a heterogenous population of tumor and normal cells

11/5/09
Variant Variants

• What if deletion boundaries are uncertain?
Observed variation in deletion size

Sizes of homozygous deletions in cell lines from different human cancers.
(scale is in megabases).
PAMP (500Kb region)

- 10 sets of 25 primers: upstream and downstream
  - 250 upstream
  - 250 downstream

- Primer-pairs closest to breakpoint amplified

- Assay by oligo array

Goal: Computational selection of an ‘optimal’ primer set
Show me the computation

- Primers are short DNA pieces (substrings of the DNA string) added to the mix to amplify specific DNA.
- We design them computationally and synthesize them in the lab
- Can we design the primers so that the reaction fails in very few patients.
- False negatives are worse than false positives!
- Q: What are the criteria that primers must satisfy so that PAMP succeeds?
The primers must be physicochemically appropriate.

Some software already selects for such primers.

Approximately 5000 substrings of the 500K region are appropriate.

We need only 500 out of the 5000
Candidate primers must satisfy:

- Melting temperature
- GC content
- Uniqueness in the genome
- Etc.

Well studied problem, and we can use standard tools to generate a candidate primer set (about 5000 primers in a 500K region).
Design criterion: coverage

- **Coverage**
  - Let $2d$ be the maximum size of an amplifiable product.
  - If the distance between two forward (reverse) primers exceeds $d$, some deletion boundaries will not be amplified.

- **We measure** coverage cost **as** $D-d$. This cost must be minimized.
Design Criterion: Dimerization

- Primers that cross-dimerize will
  1. Not participate in genomic amplification
  2. Consume PCR reagent
  3. Are ubiquitous: A candidate set of 5000 primers will have ~30000 dimerizing pairs

```
3'CTCAACAGGTGTTACATGACCTAGCACTGACCCTG-
                                       |||||           |||||   
5'-GTTTCCCAGTCACGATCGAGTTGTCCACAATGTACTG-3'

3' - GGTGTAAGACGCGAACCCTATCTAGCACT...
     |||||           
5' - GTTTCCCAGTCACGATCCATTTCCGCCACCCACATT-3'

3' - GGTTACACCTACTCTGATT-5'
     ||||x          
5' - ACAAACCCACGTCGGAAG-3'
```
Goal

• Given 5000 substrings, find 500 s.t.
  – No pair dimerizes
  – Adjacent primers are not too far apart. Minimize the cost of being far apart

• Can we test all possibilities?
Problem Formulation

- Construct a primer-coverage graph $G$
  - Each candidate primer defines a vertex, $u$, with its genomic location denoted by $l_u$. 

Vertex

<table>
<thead>
<tr>
<th>u</th>
<th>v</th>
</tr>
</thead>
</table>

Genomic Position

<table>
<thead>
<tr>
<th>$l_u$</th>
<th>$l_v$</th>
</tr>
</thead>
</table>

11/5/09
Problem Formulation

• Construct a primer-coverage graph $G$
  – The vertices are paired up with cross-hybridization edges $E_c$.
  • $(u, v) \in E_c$ if and only if primers $u$ and $v$ cross-hybridize, and co-occur in a multiplex PCR reaction.
Problem Formulation

- Construct a primer-coverage graph $G$
  - A solution to the primer design problem is a chain $P = p_1, p_2, \ldots$ of forward primers followed by reverse primers, ordered so that $l_{p_j} < l_{p_{j+1}}$ for all $j$.
Simulated annealing

- Generic technique for complex optimization problems.
Simulated annealing

- From current solution, we move to a new solution with probability $\sim \exp(-\Delta_C)$
- Non favorable solutions are also considered with low probability
S.A: Computing Neighborhoods

- Start with a current solution \(\{1,2,4,5,7,10,11\}\), with cost proportional to uncovered region.
- A neighboring solution is obtained by adding a primer, and deleting all its dimerizing partners.
Computing Neighborhoods

• A move to a neighboring solution entails a change in cost $\Delta S$
  - $\{1,2,4,5,7,10,11\} \rightarrow \{1,4,6,7,8,10,11\}$

• We move to a neighboring solution with probability proportional to $\exp(-\Delta S)$
Computing Neighborhoods

- Add \{3,6\}, remove \{9\} to improve coverage
Computing Neighborhoods

• After many iterations (convergence), we should have traversed an optimal (low-cost) solution with high probability.
Results: the CDKN2A region

• The best solution corresponds to 17,846 bp not covered (> 97% coverage)
• Another solution (better than the naïve greedy) corresponds to 103,857 bp not covered (< 80% coverage)
Coverage Plot

The graph shows the percent coverage against reverse and forward sequence size (Kb) on a log scale. The x-axis represents the reverse sequence size in multiples of 10^5, and the y-axis represents the percent coverage. The plot indicates the distribution of coverage across different sequence sizes, with distinct color bands representing different coverage levels.
A generic strategy

1. Deliver only the associated microreactors to the clinical site.
2. The personalized microreactors for monitoring of disease progression and response.

1. Sample sent to central lab
2. Microreactors mediated PAMP
3. Identification of breakpoints and associated microreactors (eg. by tiling array)
4. Confirmation of the breakpoint with associated microreactors
Human Genome, 2001
jobs: we’re hiring!

Statistical Geneticist

Location: Mountain View, CA

Job Code: 004

# of openings: 1

Description

Statistical Geneticist

Do you love data? What would you do with six billion genomes? If you have good ideas and can make them work then we want to talk to you. At 23andMe, we believe that the current crop of genetic association studies is just the beginning of a tidal wave of development in statistical genetics. This person will be involved in the analysis of human genetic data and the development of product features that depend on an intricate knowledge of both statistics and genetics. The scope and breadth of our vision means that most of these techniques have yet to be developed anywhere in the world.

Required Qualifications

- PhD in Statistical Genetics, Bioinformatics or Computer Science.
- Strong bioinformatics background; ideally AI, statistics and genetics-focused research.
- Proficiency with scripting languages.
- Background in algorithm development with C/C++ a plus.
- Familiarity with web technologies (Javascript, CSS/HTML, etc) a plus.
The Bioinformatics major

http://cse.ucsd.edu/undergrad/degreeprograms/bioinformatics.html

Contact:

- ugradinfo@cs.ucsd.edu
- (858)534-8872
- CSE Bioinformatics Committee
  - Dr. Pavel Pevzner
  - Dr. Vineet Bafna
  - Dr. Mohan Paturi