Graph Algorithms in Bioinformatics

Computational Biology
IST
Ana Teresa Freitas
2015/2016

Sequencing

Clone-by-clone shotgun sequencing
Human Genome Project

Whole-genome shotgun sequencing
Celera Genomics

(BACs)
Assembling Genomes

- Must take the fragments and put them back together
  - Not as easy as it sounds.
- SCS Problem (Shortest Common Superstring)
  - Fit overlapping sequences together to get the shortest possible sequence that includes all fragment sequences
- DNA fragments contain sequencing errors
- Two complements of DNA
  - Need to take into account both directions of DNA
- Repeat problem
  - 50% of human DNA is just repeats
  - If you have repeating DNA, how do you know where it goes?

DNA Sequencing: History

- Sanger sequencing method
  - Fred Sanger et al (1977)
  - Method takes advantage of how cells make copies of DNA
  - Run one starvation experiment for each of A, T, G and C
  - Separate the resulting DNA fragments by length (Sanger ladder) (electrophoresis)
DNA Sequencing: History (cont’d)

- Maxam – Gilbert sequencing method
  - Walter Gilbert et al (1977)
  - Chemical method to cleave DNA at specific points (G, G +A, T+C, C).
  - Labeled fragments of varying lengths generated & electrophoresed.

DNA Sequencing

- Shear DNA into millions of small fragments
- Read 500 – 700 nucleotides at a time from the small fragments (Sanger method)
Fragment Assembly

• **Computational Challenge**

Fragment assembly – compose a single sequence, a “superstring”, by assembling individual reads

---

Shortest Superstring Problem

• **Problem:** Given a set of strings, find a shortest string that contains all of them
• **Input:** Strings $s_1, s_2, \ldots, s_n$
• **Output:** A string $s$ that contains all strings $s_1, s_2, \ldots, s_n$ as substrings, such that the length of $s$ is minimized

• **Complexity:** NP – complete
• **Note:** this formulation does not take into account sequencing errors & repeats
Shortest Superstring Problem: Example

The Shortest Superstring problem

Set of strings: \{000, 001, 010, 011, 100, 101, 110, 111\}

Concatenation
Superstring 000 001 010 011 100 101 110 111

The shortest superstring in this case represents a solution of the Clever Thief problem.

It is the minimum string of tests a thief has to conduct to try all possible k-letter passwords for a combination lock.

Reducing SSP to TSP

- Define overlap \((s_i, s_j)\) as the length of the longest prefix of \(s_j\) that matches a suffix of \(s_i\).

- Construct a graph with \(n\) vertices representing the \(n\) strings \(s_1, s_2, \ldots, s_n\). Insert edges of length overlap \((s_i, s_j)\) between vertices \(s_i\) and \(s_j\).

- Find the shortest path which visits every vertex exactly once. This is the Traveling Salesman Problem (TSP), which is also NP – complete.
Reducing SSP to TSP (cont’d)

SSP to TSP: An Example

\[ S = \{ \text{ATC}, \text{CCA}, \text{CAG}, \text{TCC}, \text{AGT} \} \]

**SSP**

AGT  
CCA  
ATC  
**ATCCAGT**  
TCC  
CAG

**TSP**

AGT  
ATC  
CCA  
**ATCCAGT**  
TCC  
CAG  

0 2 1 1 2 1 2
Sequencing by Hybridization (SBH): History

- **1988**: SBH suggested as alternative sequencing method. Nobody believed it will ever work
- **1991**: Light directed polymer synthesis developed by Steve Fodor and colleagues.
- **1994**: Affymetrix develops first 64-kb DNA microarray

Microarrays

- Rows represent genes
- Columns represent samples
**DNA polymer: the double helix**

- [Illustration of DNA structure]

**DNA chips**

- **Solid flat surface** containing probe DNA molecules in a matrix-like pattern (array), to detect complementary DNA strands

- [Diagram of DNA chip]

- **Biological Sample** → **Biological Information**
  - • Gene Expression
  - • Disease Diagnosis
  - • Drug Discovery
  - • Sequencing
DNA chips - detection

- **Target DNA is labeled with fluorescent molecules.**
- **Detection of hybridization by laser scanning and imaging.**

![DNA chips image](Incyte, www.incyte.com)

---

How SBH Works

- Attach all possible DNA probes of length $l$ (e.g., $l = 8$) to a flat surface, each probe at a distinct and known location. This set of probes is called the DNA array.
- Apply a solution containing fluorescently labeled DNA fragment to the array.
- The DNA fragment hybridizes with those probes that are complementary to substrings of length $l$ of the fragment.
How SBH Works (cont’d)

• Using a spectroscopic detector, determine which probes hybridize to the DNA fragment to obtain the $l$–mer composition of the target DNA fragment.

• Apply the combinatorial algorithm described below to reconstruct the sequence of the target DNA fragment from the $l$–mer composition.

Hybridization on DNA Array
**l-mer composition**

- Define $Spectrum(s, l)$ as the unordered multiset of all possible $(n - l + 1)\ l$-mers in a string $s$ of length $n$

- The order of individual elements in $Spectrum(s, l)$ does not matter

---

**l-mer composition**

- For example, for $s = TATGGTGC$ all of the following are equivalent representations of $Spectrum(s, 3)$:
  - \{TAT, ATG, TGG, GGT, GTG, TGC\}
  - \{ATG, GGT, GTG, TAT, TGC, TGG\}
  - \{TGG, TGC, TAT, GTG, GGT, ATG\}
\-mer composition

- For example, for \( s = \text{TATGGTGC} \) all of the following are equivalent representations of \( \text{Spectrum}(s, 3) \):
  - \{\text{TAT, ATG, TGG, GGT, GTG, TGC}\}
  - \{\text{ATG, GGT, GTG, TAT, TGC, TGG}\}
  - \{\text{TGG, TGC, TAT, GTG, GGT, ATG}\}

  We usually choose the lexicographically maximal representation as the canonical one.

Different sequences – the same spectrum

- Different sequences may have the same spectrum:
  - \( \text{Spectrum(GTATCT,2)} = \text{Spectrum(GTCTAT,2)} = \{\text{AT, CT, GT, TA, TC}\} \)
The SBH Problem

• **Goal**: Reconstruct a string from its $l$-mer composition

• **Input**: A set $S$, representing all $l$-mers from an (unknown) string $s$

• **Output**: String $s$ such that $Spectrum(s, l) = S$

SBH: Hamiltonian Path Approach

$S = \{ \text{ATG, AGG, TGC, TCC, GTC, GGT, GCA, CAG} \}$

Path visited every VERTEX once
SBH: Hamiltonian Path Approach

\[ S = \{ \text{ATG, TGG, TGC, GTG, GGC, GCA, GCG, CGT} \} \]

- This spectrum yields a more complicated graph with two Hamilton paths.

- As the overlap graph becomes larger, this approach ceases to be practically useful since the Hamiltonian Path problem is NP-complete.

SBH: Eulerian Path Approach

\[ S = \{ \text{ATG, TGG, TGC, GTG, GGC, GCA, GCG, CGT} \} \]

Vertices correspond to \((l-1)-mers\): \{ AT, TG, GC, GG, GT, CA, CG \}

Edges correspond to \(l\)-tuples from \(S\)

Path visited every EDGE once
SBH: Eulerian Path Approach

S = \{ AT, TG, GC, GG, GT, CA, CG \} corresponds to two different paths:

\[ \text{ATGGCGTGCA} \quad \text{ATGCGTGGCA} \]

Balanced Graphs

- A graph is balanced if for every vertex the number of incoming edges equals to the number of outgoing vertices:
  \[ \text{in}(v) = \text{out}(v) \]
Euler Theorem

- A graph is balanced if for every vertex the number of incoming edges equals to the number of outgoing vertices:
  \[ \text{in}(v) = \text{out}(v) \]
- **Theorem**: A connected graph is Eulerian if and only if each of its vertices is balanced.

Euler Theorem: Extension

- **Theorem**: A connected graph has an Eulerian path if and only if it contains at most two semi-balanced vertices and all other vertices are balanced.
Some Difficulties with SBH

- **Fidelity of Hybridization**: difficult to detect differences between probes hybridized with perfect matches and 1 or 2 mismatches
- **Array Size**: Effect of low fidelity can be decreased with longer \( l \)-mers, but array size increases exponentially in \( l \). Array size is limited with current technology.
- **Practicality**: SBH is still impractical. As DNA microarray technology improves, SBH may become practical in the future

Shotgun Sequencing

Get one or two reads from each segment

~500 bp

 genomic segment

cut many times at random (Shotgun)
**Fragment Assembly**

Cover region with ~7-fold redundancy
Overlap reads and extend to reconstruct the original genomic region

**Read Coverage**

Length of genomic segment: $L$
Number of reads: $n$
Length of each read: $l$

**Definition:** Coverage $C = \frac{n \cdot l}{L}$
Enough Coverage

How much coverage is enough?

**Lander-Waterman model:**
Assuming uniform distribution of reads, $C=10$ results in 1 gapped region per 1,000,000 nucleotides

---

Challenges in Fragment Assembly

- Repeats: A **major** problem for fragment assembly
- > 50% of human genome are repeats:
  - over 1 million *Alu* repeats (about 300 bp)
  - about 200,000 LINE repeats (1000 bp and longer)

Green and blue fragments are interchangeable when assembling repetitive DNA
Repeat Types

- **Low-Complexity DNA** (e.g. ATATATACATA...)
- **Microsatellite repeats** \((a_1...a_k)^n\) where \(k \approx 3-6\) (e.g. CAGCAGTAGCAGCACCAG)
- **Transposons/retrotransposons**
  - **SINE** Short Interspersed Nuclear Elements (e.g., **Alu**: \(~300\) bp long, \(10^6\) copies)
  - **LINE** Long Interspersed Nuclear Elements \(~500 - 5,000\) bp long, \(200,000\) copies
  - **LTR retroposons** Long Terminal Repeats \(~700\) bp at each end
- **Gene Families** genes duplicate & then diverge
- **Segmental duplications** \(~very\ long,\ very\ similar\ copies\)

Example

- Try to solve the **SBH** problem for the following spectrum:

  \[ S = \{ATG, GGG, GGT, GTA, GTG, TAT, TGG\} \]

  - First, build the graph representing edges and vertexes
  - Identify all possible sequences \(s\), such that \(\text{Spectrum}(s,3) = S\)