Algorithmic Approaches for Biological Data, Lecture #18

Katherine St. John

City University of New York
American Museum of Natural History

13 April 2016

Outline

- Comparing & Aligning Sequences

Outline

- Comparing & Aligning Sequences
- Longest Common Substrings

Outline

- Comparing & Aligning Sequences
- Longest Common Substrings
- Dynamic Programming Example: Manhattan Tourist Problem

Comparing & Aligning sequences

A C G T C C T C

A C G C C T A C

Comparing & Aligning sequences

A C G T C C T C

A C G C C T A C

- How do you compare two sequences?

Comparing & Aligning sequences

A C G T C C T C

A C G C C T A C

- How do you compare two sequences?
- Hamming Distance: count the pairwise differences:

A C G T C C T C

A C G C C T A C

Comparing & Aligning sequences

A C G T C C T C

A C G C C T A C

Comparing & Aligning sequences

A C G T C C T C

A C G C C T A C

- Where do the differences come from?

Comparing & Aligning sequences

A C G T C C T C

A C G C C T A C

- Where do the differences come from?
- If point mutations, Hamming Distance captures this well.

Comparing & Aligning sequences

A C G T C C T C

A C G C C T A C

- Where do the differences come from?
- If point mutations, Hamming Distance captures this well.
- What if the changes come from insertions or deletions into the sequence?

Comparing & Aligning sequences

A C G T C C T C

A C G C C T A C

- Where do the differences come from?
- If point mutations, Hamming Distance captures this well.
- What if the changes come from insertions or deletions into the sequence?
- Then would expect missing sections (gaps) and should align the sequences.

Comparing & Aligning sequences

```
A C G T C C T - C
A C G - C C T A C
```

- Where are the similarities?

Comparing & Aligning sequences

```
A C G T C C T - C
A C G - C C T A C
```

- Where are the similarities?
- What is the longest common subsequence (gaps allowed)?

Comparing & Aligning sequences

```
A C G T C C T - C
A C G - C C T A C
```

- Where are the similarities?
- What is the longest common subsequence (gaps allowed)?
- Easy to identify on the aligned sequence.

Comparing & Aligning sequences

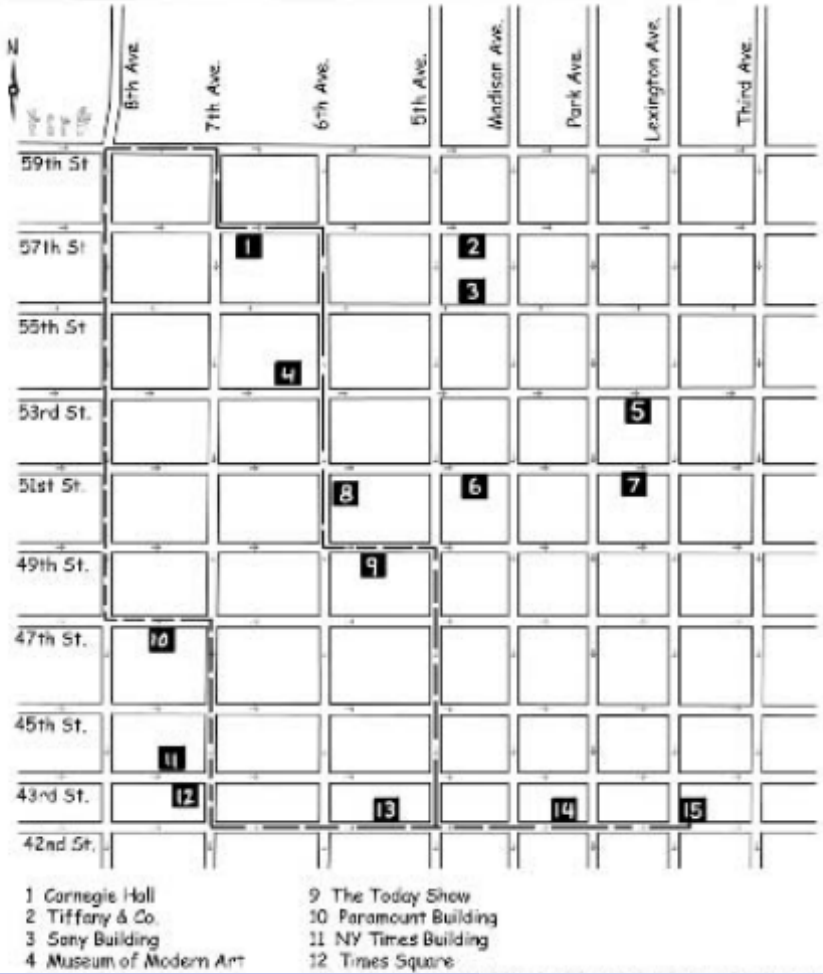
```
A C G T C C T - C
A C G - C C T A C
```

- Where are the similarities?
- What is the longest common subsequence (gaps allowed)?
- Easy to identify on the aligned sequence.

In Pairs

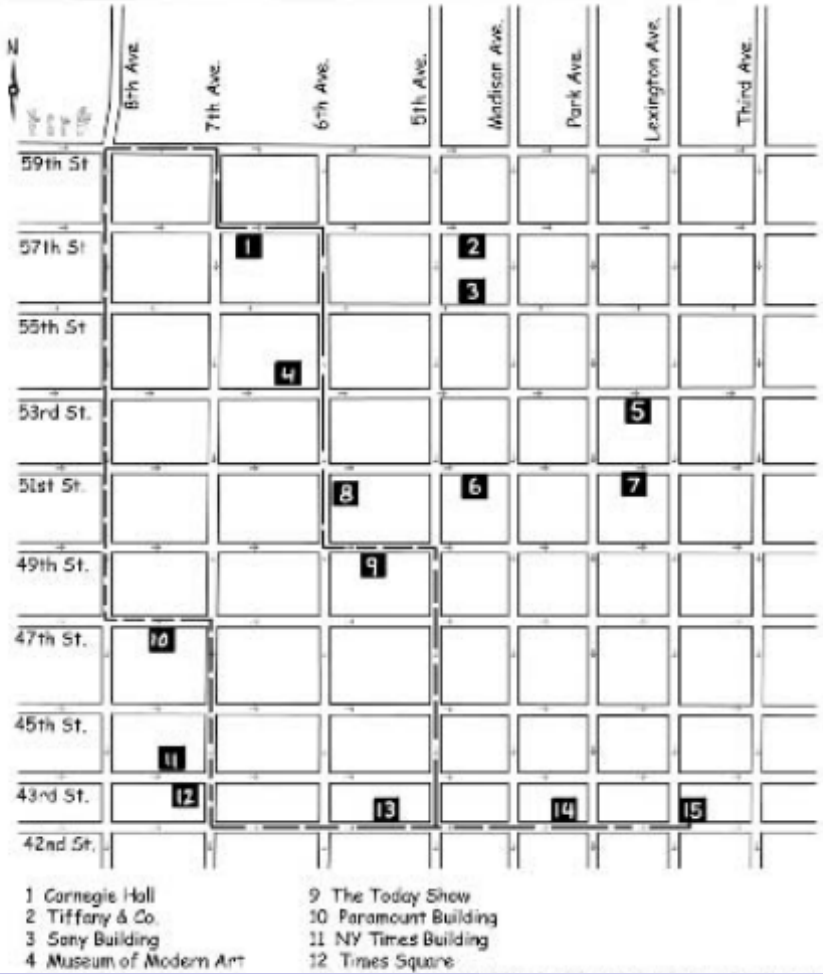
- Given the sequences: TATATATAAAAA
  ATGATGAAAAAAAA
  - What is the Hamming distance of the sequences?
  - Can you lower the distance by allowing gaps?
- Given the sequences: ATGCAATTCAGTTCCGTGGACTACATGGTCTACTTTCAG
  TGCAAAATTCAGTTCCGTGGACTACATGGGTCTACTTCAG
  - What is the Hamming distance of the sequences?
  - Can you lower the distance by allowing gaps?
- What parts can you automate of this process? Sketch an algorithm.
- Manhattan Tourist Problem (next slide and handout).

In Pairs: Manhattan Tourist Problem



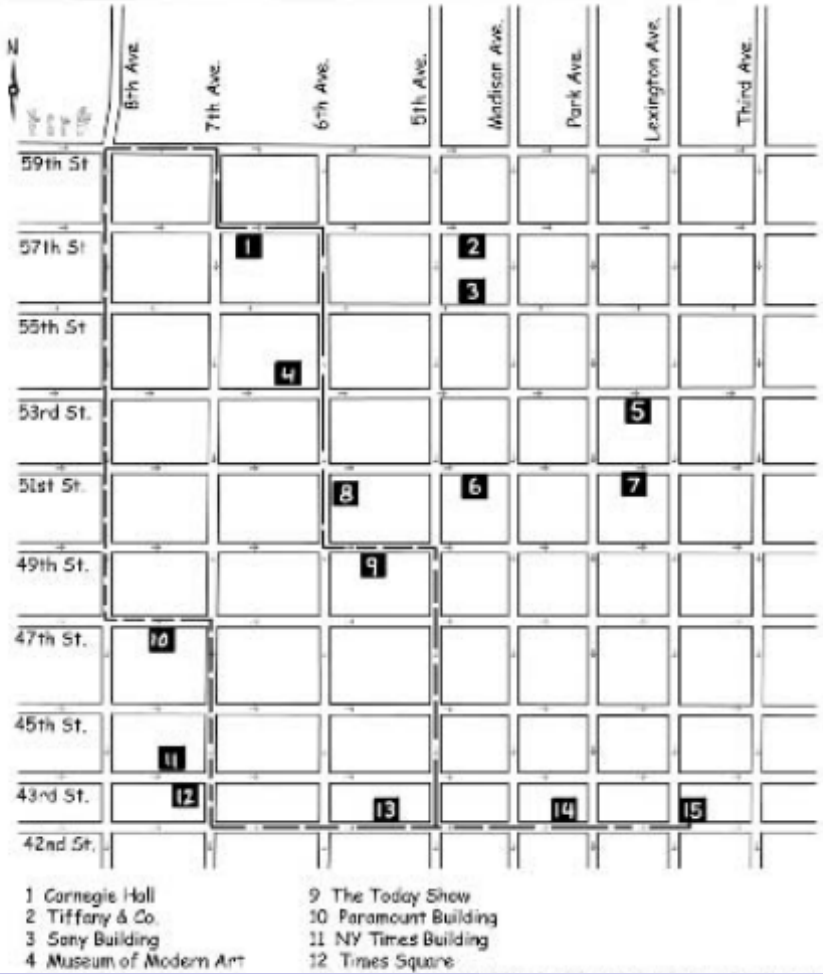
- In hurry, and want to visit as many landmarks as possible.

In Pairs: Manhattan Tourist Problem



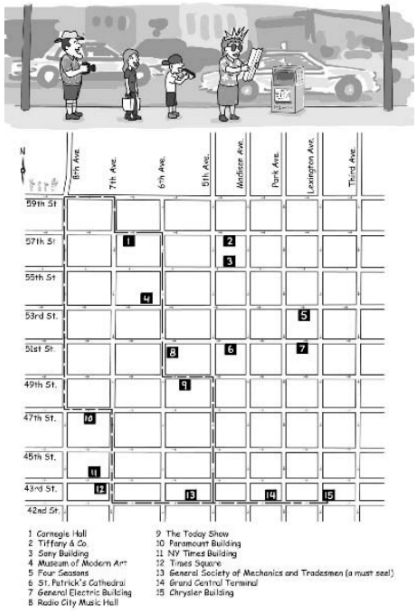
- In hurry, and want to visit as many landmarks as possible.
- Can only walk south and east.

In Pairs: Manhattan Tourist Problem



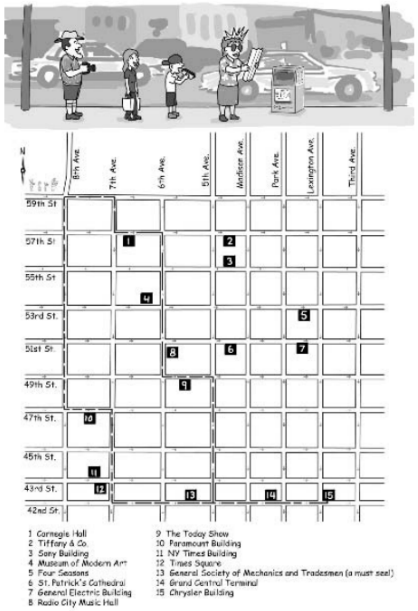
- In hurry, and want to visit as many landmarks as possible.
- Can only walk south and east.
- What's the best route?

Dynamic Programming



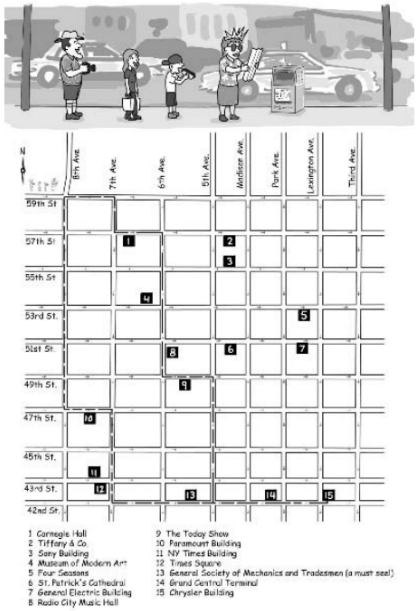
- Both questions have an interesting property:

Dynamic Programming



- Both questions have an interesting property:
  - When searching for the optimal answer, the same subproblems occur multiple times.

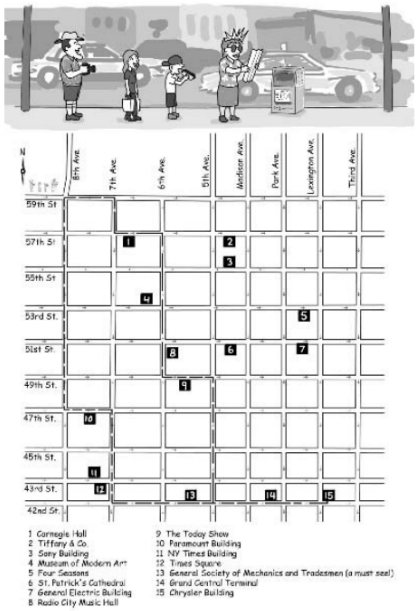
Dynamic Programming



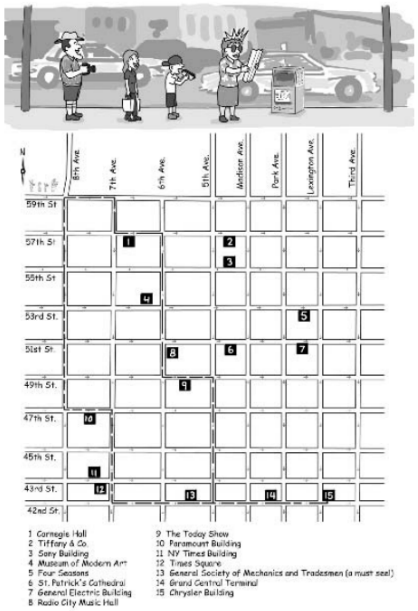
- Both questions have an interesting property:
  - When searching for the optimal answer, the same subproblems occur multiple times.
- Instead of re-computing each time, store in a table to be re-used later.

Dynamic Programming

- Instead, store computations in a table to be re-used:

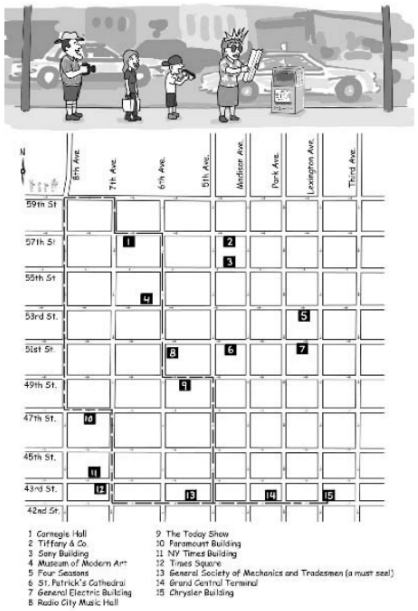


Dynamic Programming



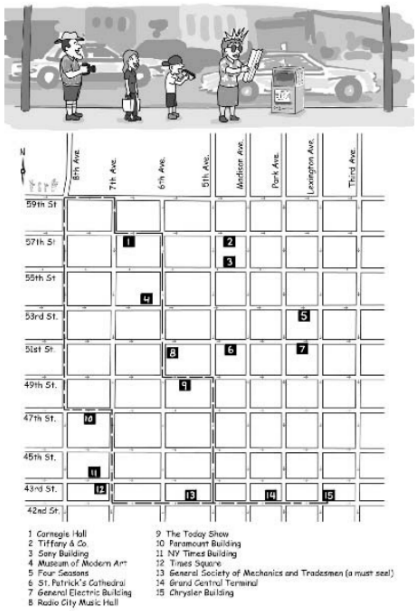
- Instead, store computations in a table to be re-used:
- For example, for the tourists: working backwards:

Dynamic Programming



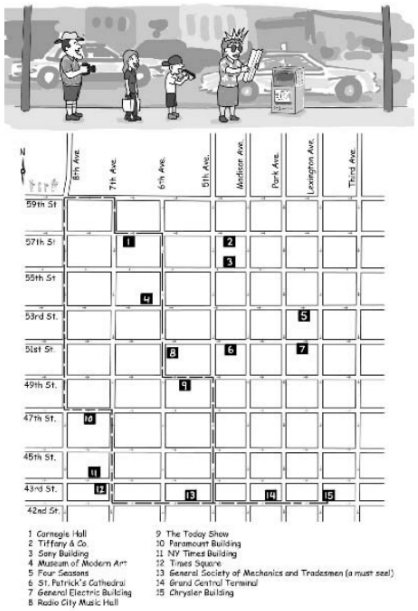
- Instead, store computations in a table to be re-used:
- For example, for the tourists: working backwards:
- The optimal number will be the best of: arriving from 42 & Lex or from 43 & Third.

Dynamic Programming



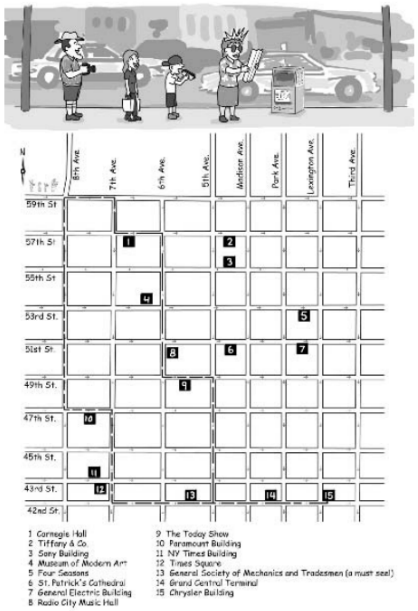
- Instead, store computations in a table to be re-used:
- For example, for the tourists: working backwards:
- The optimal number will be the best of: arriving from 42 & Lex or from 43 & Third.
  - From 42 & Lex, add 1 (Chrysler Building) to best from either 42 & Park or 43 & Lex.

Dynamic Programming



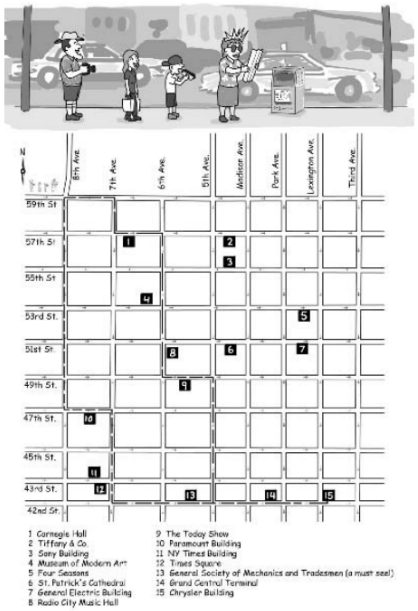
- Instead, store computations in a table to be re-used:
- For example, for the tourists: working backwards:
- The optimal number will be the best of: arriving from 42 & Lex or from 43 & Third.
  - From 42 & Lex, add 1 (Chrysler Building) to best from either 42 & Park or 43 & Lex.
  - From 43 & Third, choose to best from either 43 & Lex and 45 & Third.

Dynamic Programming



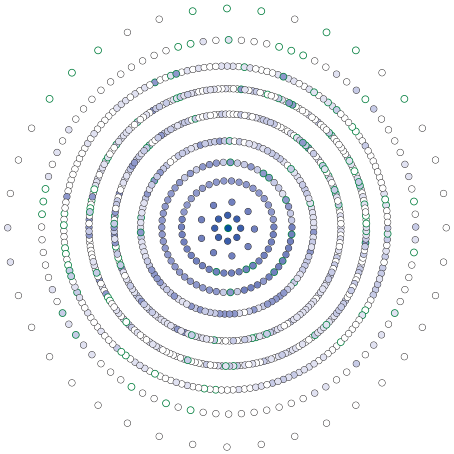
- Instead, store computations in a table to be re-used:
- For example, for the tourists: working backwards:
- The optimal number will be the best of: arriving from 42 & Lex or from 43 & Third.
  - From 42 & Lex, add 1 (Chrysler Building) to best from either 42 & Park or 43 & Lex.
  - From 43 & Third, choose to best from either 43 & Lex and 45 & Third.
- 43 & Lex is used in both options– store it's value in a table so it only has to be computed once.

Dynamic Programming



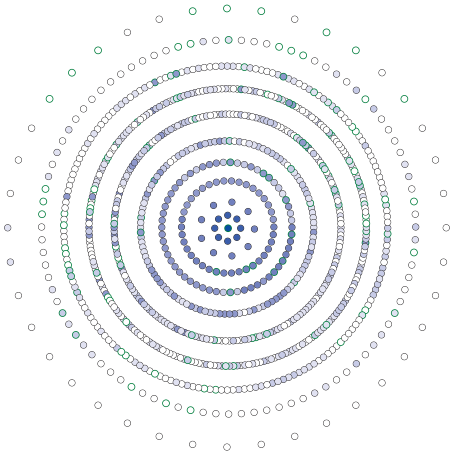
- Instead, store computations in a table to be re-used:
- For example, for the tourists: working backwards:
- The optimal number will be the best of: arriving from 42 & Lex or from 43 & Third.
  - From 42 & Lex, add 1 (Chrysler Building) to best from either 42 & Park or 43 & Lex.
  - From 43 & Third, choose to best from either 43 & Lex and 45 & Third.
- 43 & Lex is used in both options– store it's value in a table so it only has to be computed once.
- The approach of computing answers to the subproblems for later use in the optimization is called dynamic programming. (Much more on this next time.)

Recap



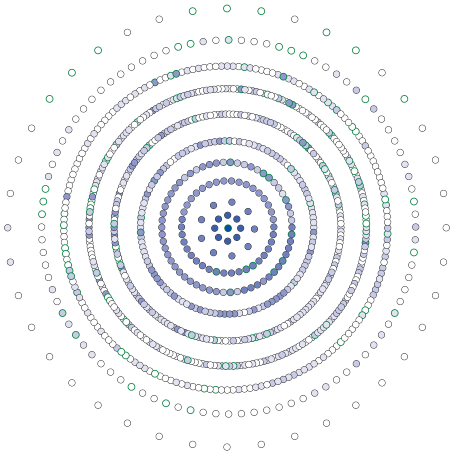
- Using `sqlitebrowser` in lab today (SQL & Databases).

Recap



- Using `sqlitebrowser` in lab today (SQL & Databases).
- Email lab reports to `kstjohn@amnh.org`

Recap



- Using `sqlitebrowser` in lab today (SQL & Databases).
- Email lab reports to `kstjohn@amnh.org`
- Challenges available at `rosalind.info`