CSCI 474
Bioinformatics

Filip Jagodzinski
Announcement - Article 3 summary and critique

- Three articles will be available. Will be posted to the course website by Tuesday.
- You’ll read one, and compose a summary and critique.
- You’ll then review other students’ summaries and critiques who read an article that is different from the one that you read.
Announcement- Article 3 summary and critique

• Three articles will be available. Will be posted to the course website by Tuesday

• You’ll read one, and compose a summary and critique.

• You’ll then review other students’ summaries and critiques who read an article that is different from the one that you read

• ISB trip interest
From last time ...

Assume you have the following 4 sequences. Align them ALL such that they are “best” aligned.

A G C T
A G A G C T
A G G G T
A G T G
From last time ...

Assume you have the following 4 sequences. Align them ALL such that they are “best” aligned.

A G C T
A G A G C T
A G G G T
A G T G

62

A G C T
A G A G C T
A G G G T
A G T G

40

55

65
From last time ...

**Scoring**
- **Match** = +1
- **Mismatch** = -1
- (internal or external) **Gap** = -2

**Original sequences**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>T</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>T</td>
<td>G</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>T</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>T</td>
<td>G</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>T</td>
<td>G</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scoring</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Match</td>
<td>+1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mismatch</td>
<td></td>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>Gap</td>
<td></td>
<td></td>
<td>-2</td>
</tr>
</tbody>
</table>

**Goal**: Perform an optimal alignment such that the alignment “score” is optimized.

---

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>-2</td>
<td>--</td>
<td>0</td>
</tr>
<tr>
<td>G-</td>
<td>-2</td>
<td>-G</td>
<td>-2</td>
</tr>
<tr>
<td>GG</td>
<td>+1</td>
<td>-G</td>
<td>-2</td>
</tr>
<tr>
<td>GG</td>
<td>+1</td>
<td></td>
<td>-4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-2</td>
<td>GG</td>
</tr>
</tbody>
</table>
From last time ...

**Scoring**

- **Match** = +1
- **Mismatch** = -1

(Internal or external) **Gap** = -2

Original sequences

<table>
<thead>
<tr>
<th>A</th>
<th>T</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>T</td>
<td>G</td>
</tr>
<tr>
<td>A</td>
<td>T</td>
<td>G</td>
</tr>
</tbody>
</table>

+2 0 -9

-2 -2 +10 +6

Trying to solve the sub-problems might lead us into trouble (hence the non-optimality of the solution which uses heuristic)

<table>
<thead>
<tr>
<th>A</th>
<th>T</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>T</td>
<td>G</td>
</tr>
<tr>
<td>A</td>
<td>T</td>
<td>G</td>
</tr>
</tbody>
</table>

+2 -9 +10 +3
From last time ...

**Scoring**
- Match = +1
- Mismatch = -1
- Gap = -2

(Internal or external)

**Original sequences**

```
ATG
AGT
ATG
ATG
+2 0 -9 -7
```

**Q: What are the subproblem scores?**

Score:

```
ATG
AG
TG
ATG
ATG
-2 -2+10 +6
```

Score:

```
ATG
AG
TG
ATG
ATG
+2 -9+10 +3
```
From last time …

Scoring
Match = +1
Mismatch = -1
(internal or external) Gap = -2

Original sequences

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>T</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>A</td>
<td>T</td>
<td>G</td>
</tr>
<tr>
<td>1</td>
<td>A</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>T</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>T</td>
<td>G</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>T</td>
<td>G</td>
</tr>
</tbody>
</table>

Score:
-2 \( -2 + 1 = -3 \)

Q: What are the subproblem scores?

The green sub-problem has a “better” score than the blue sub-problem, but it is part of the LOWER overall scores.

Score: \(-1 + 0 + 1 = 0\)
From last time ...

### Scoring
- **Match** = +1
- **Mismatch** = -1
- **Gap** = -2

Original sequences:

<table>
<thead>
<tr>
<th>T</th>
<th>C</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>G</td>
<td>T</td>
<td>C</td>
</tr>
</tbody>
</table>

### Progressive Alignment

1. Align all sets of pairwise sequences. First align TCG with CGT, and CGT with GTC, and TCG with GTC.
2. For each sub-problem pair, align to it the “left” out sequence, and do NOT move the already aligned sub-problem.
From last time ...

**Scoring**
- Match = +1
- Mismatch = -1
- (internal or external) Gap = -2

Original sequences

```
T C G
C G T
G T C
```

Q: Is there any hope?

Q: Ideas on algorithms?

```
T C G
C G T
G T C
```

Define all possible 2-sequence sub-problems
From last time ...

Scoring
Match = +1
Mismatch = -1
(internal or external) Gap = -2

Original sequences

Q: Is there any hope?
Q: Ideas on algorithms?

Optimally align the 3 sets
From last time …

Scoring
Match = +1
Mismatch = -1
(internal or external) Gap = -2

Q: Is there any hope?
Q: Ideas on algorithms?

Original sequences

Into each, align the “left out” sequence without modifying the already solved sub-problem
From last time ...

**Scoring**
- Match = +1
- Mismatch = -1
- (internal or external) Gap = -2

**Q: Is there any hope?**

**Q: Ideas on algorithms?**

Original sequences

Into each, align the “left out” sequence without modifying the already solved sub-problem.

Scoring:
- Match = +1
- Mismatch = -1
- Gap = -2
From last time ...

**Scoring**
- Match = +1
- Mismatch = -1
- Gap = -2

*(internal or external)*

Original sequences

<table>
<thead>
<tr>
<th>T</th>
<th>C</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>G</td>
<td>T</td>
<td>C</td>
</tr>
</tbody>
</table>

Q: Is there any hope?

Q: Ideas on algorithms?

Into each, align the “left out” sequence without modifying the already solved sub-problem
Today

Structural Bioinformatics
Mutations
Protein structure
Q: What is a mutation?
Q: What are the biophysical causes of Huntington’s Disease?
Motivation – Huntington’s disease

The *Huntington* gene is located on 4p16.3

Q: Where in the human genome is the HTT gene?
Motivation – Huntington’s disease

The *Huntington* gene is located on 4p16.3

Tri-nucleotide repeat
CAG : Glutamine

...CAGCAGCAG...
Motivation – Huntington’s disease

Tri-nucleotide repeat
CAG : Glutamine

...CAGCAGCAG...
Motivation – Huntington’s disease

Tri-nucleotide repeat
CAG : Glutamine

...CAGCAGCAG...

Repeats: fewer than 26  Normal
27-35             Intermediate, no physiological effects
36-39             May have physiological effects
40+                Physiological effects
Motivation – Huntington’s disease

Tri-nucleotide repeat
CAG : Glutamine

...CAGCAGCAG...

Nuclear inclusions
Motivation – Huntington’s disease

DNA -> Tri-nucleotide repeats -> malformed protein -> disrupts neuron function

Tri-nucleotide repeat
CAG : Glutamine

. . . CAGCAGCAG . . .

Nuclear inclusions

Q: What details about the protein’s structure gives insights about the effect of the HTT mutation?
Protein structure

Q: What does the period table tell us about how atoms interact?
Q: What does the period table tell us about how atoms interact?

![Periodic Table](image)
Q: These are what two amino acids?
Protein structure

Twenty-One Amino Acids

A. Amino Acids with Electrically Charged Side Chains

- Arginine (Arg)
- Histidine (His)
- Lysine (Lys)
- Arginic Acid (Asp)
- Glutamic Acid (Glu)

B. Amino Acids with Polar Uncharged Side Chains

- Serine (Ser)
- Threonine (Thr)
- Asparagine (Asn)
- Glutamine (Gln)

C. Special Cases

- Cysteine (Cys)
- Selenocysteine (Sec)
- Glycine (Gly)
- Proline (Pro)

D. Amino Acids with Hydrophobic Side Chains

- Alanine (Ala)
- Isoleucine (Ile)
- Leucine (Leu)
- Methionine (Met)
- Phenylalanine (Phe)
- Tryptophan (Trp)
- Tyrosine (Tyr)
- Valine (Val)
Protein structure

A GUIDE TO THE TWENTY COMMON AMINO ACIDS

**Chart Key:**
- **Aromatic**
- **Aliphatic**
- **Hydrophobic**
- **Polar**
- **Anionic**
- **Aromatic**
- **Non-Essential**
- **Essential**

**Amino Acids:**
- Alanine
- Valine
- Leucine
- Isoleucine
- Threonine
- Serine
- Cysteine
- Methionine
- Proline
- Hydroxyproline

**Notes:**
- This chart only shows amino acids that are found in the human genome.
- The shaded colors indicate the chemical properties of the amino acids.
- Hydrophobic amino acids are generally found in the interior of proteins.

**Essential Amino Acids:**
- Isoleucine
- Leucine
- Methionine
- Phenylalanine
- Threonine
- Tryptophan
- Valine

**Non-Essential Amino Acids:**
- Alanine
- Asparagine
- Aspartate
- Glycine
- Glutamine
- Glutamate
- Proline
- Hydroxyproline

**Other Amino Acids:**
- Cysteine
- Methionine
- Proline
- Hydroxyproline

**Special Note:**
- In some cases, distinguishing between hydroxyproline and proline can be difficult. In these cases, refer to the chart for more information.
Protein structure

Phenylalanine → Tyrosine → Tryptophan
Protein structure
Protein structure
Q: What types of constraints are imposed due to atomic identities and placement of atoms?
Protein structure

Q: Are there other constraints among atoms?
Protein structure
Protein structure
Protein structure

- Backbone Angles
- Sidechain Attraction
- Sidechain repulsion
- Bending
- Twisting
- Rigidity
- Non-twisting
Protein structure
Q: How are atoms and their positions in proteins identified?
Protein structure – X-ray crystallography
Protein structure – X-ray crystallography

2d \sin \theta = n\lambda
Protein structure – X-ray crystallography
Protein structure – X-ray crystallography
Protein structure – X-ray crystallography
Protein structure – X-ray crystallography

- X-Rays → Diffraction Pattern
- Solve for Electron Densities → Electron Density Map
- Fitting → Refinement → Fitting
Protein structure – Motivation

Q: Why not ... . . . CAGCAGCAG . . .
Protein structure – Motivation

Q: Why not ...  

...CAGCAGCAG...
Protein structure – Motivation

Q: Why not ... \( \ldots CAGCAGCAG \ldots \)
Q: Why not ... \[\ldots CAGCAGCAG \ldots\] ... Compare the structures to indirectly infer the effect of the mutation.
Mutation Analysis