Week 4 Discussion Section
Genome 540

Chengxiang Qiu

Agenda

- Q’s about HW3?
- Introduction on Github
- Overview of HW4
- Other DAG Algorithms

HW3 output summary

- Nucleotide histogram
- Background nt frequencies (based on both strands)
- Count matrix (-10 to +10 nucleotides)
- Frequency matrix (-10 to +10 nucleotides)
- Weight matrix (-10 to +10 nucleotides)
- Maximum score
- Score histogram for annotated CDS TSSs
- Score histogram for all positions
- List of non-CDS positions with score >=10

Version control - Github

Version control - GitHub

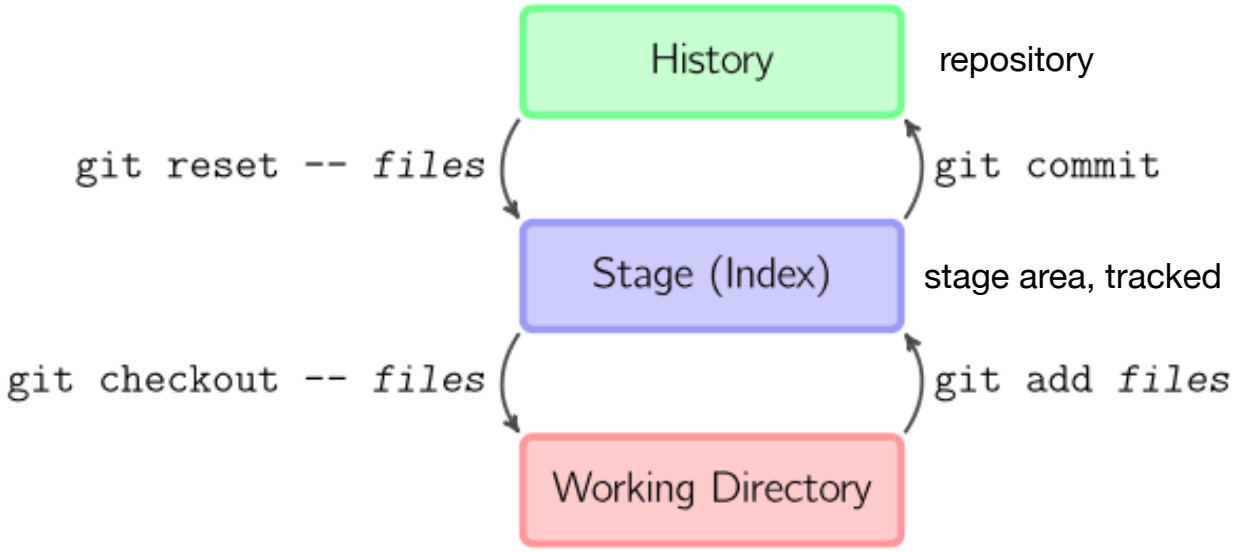
1. Easy for version control (adding new function, tracking updates, multiple platforms)

2. Easy for multiple people contributing the same project (branches and commits)

3. Easy for sharing/publishing programs (Github for code, GEO for data)

4. Build some light servers

Version control - GitHub



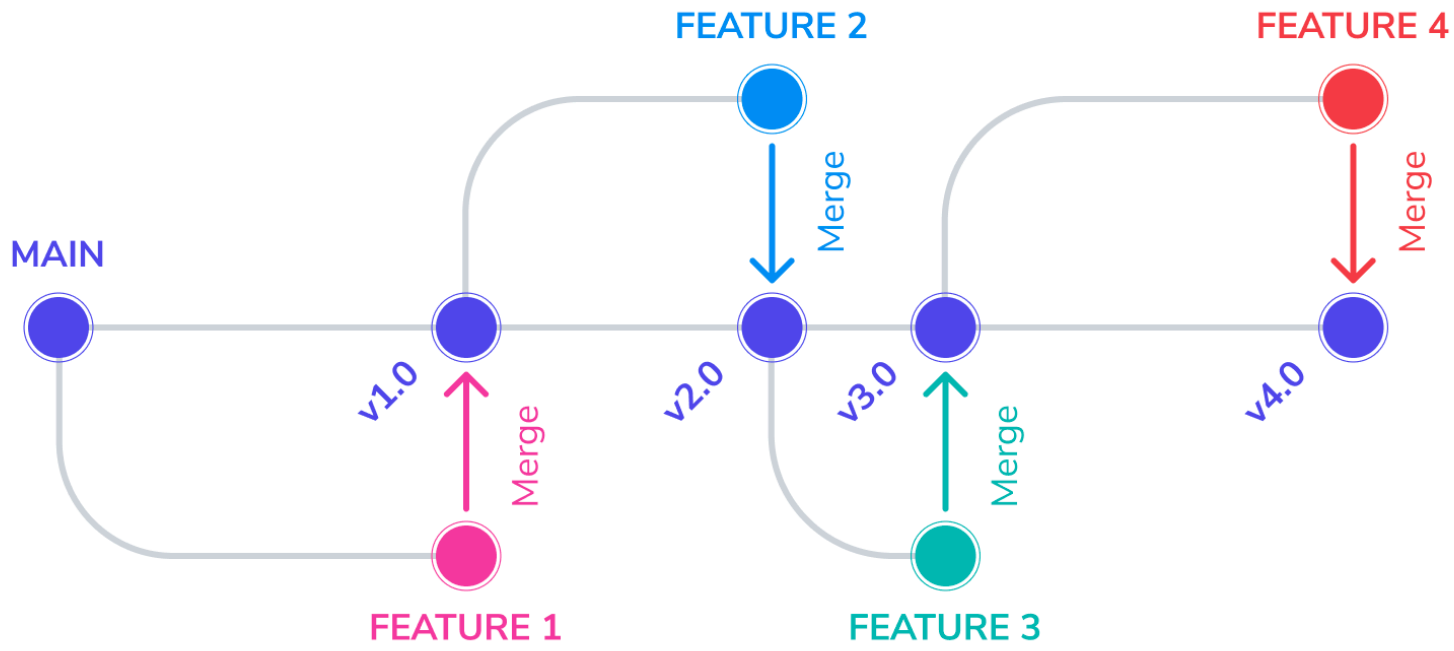
git add

git diff

git rm

git commit

Version control - GitHub



git branch

git branch NAME

git checkout

git merge

Version control - GitHub

git clone https://github.com/cole-trapnell-lab/monocle-release.git
or git clone url -b branch-name ### clone a specific branch

git branch a-new-branch ### create a new branch

git checkout a-new-branch ### switch to the new branch

git add test.file ### add new file to the stage area

git commit test.file -m "add test.file” ### add new file to repository

git push -u origin a-new-branch ### push the new branch to server

git checkout main; git pull origin master ### switch back to master branch, and update it

HW4: Highest Weighted Path for Directed Acyclic Graphs (DAGs)

- Write 2 programs
- Test your code on our examples and compare your results
- Run your code on the submission DAG and genome and submit these results

HW4: Write two programs

Program 1: Find a highest-weight path in a weighted directed acyclic graph

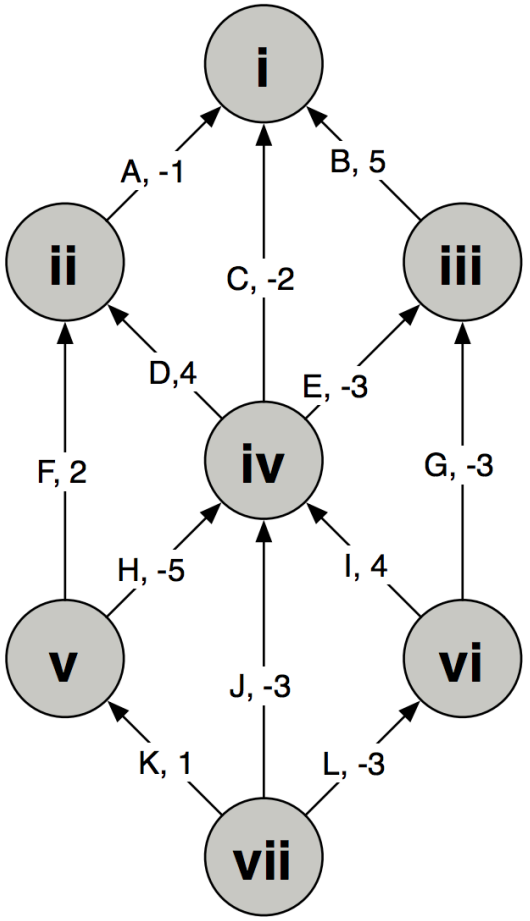
Program 2: Convert a fasta file to written graph format

HW4: Program 1

Program 1: Find a highest-weight path in a weighted directed acyclic graph

- Convert graph by hand into a list of vertices and edges, which you will input into your program
- Write a program to output the
    1. Path Score
    2. The label of the beginning vertex on the path
    3. the label of the ending vertex on the path
    4. (in order) the labels for all the edges that occur on this path.

Test Example:

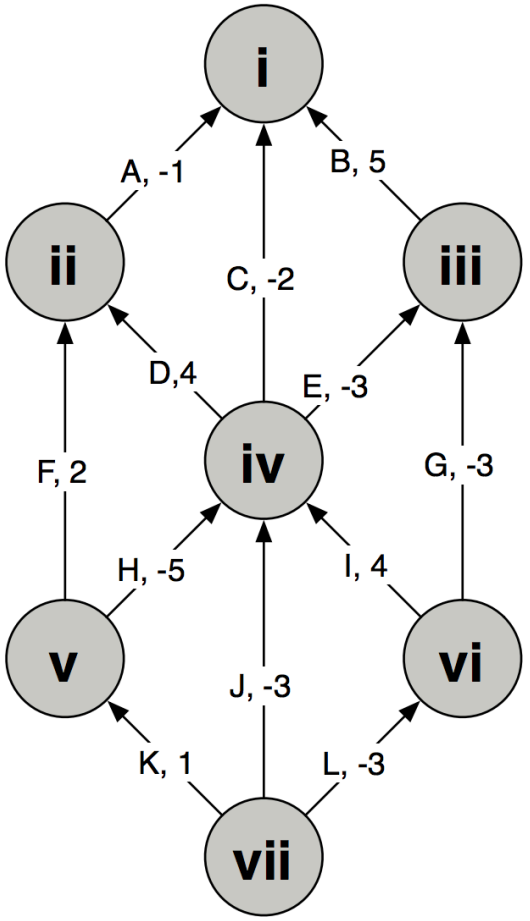


HW4: Program 1

Program 1: Find a highest-weight path in a weighted directed acyclic graph

- Convert graph by hand into a list of vertices and edges, which you will input into your program
- Write a program to output the
    1. Path Score
    2. The label of the beginning vertex on the path
    3. the label of the ending vertex on the path
    4. (in order) the labels for all the edges that occur on this path.
- Run this program “unconstrained” and with the start and end constraints of “vii” and “i” respectively

Test Example:



HW4: Program 1

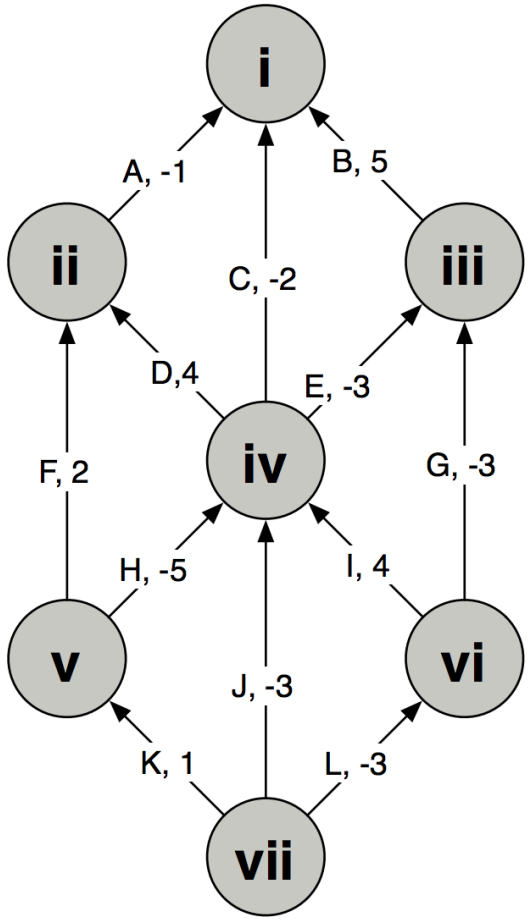
Program 1: Find a highest-weight path in a weighted directed acyclic graph

- Convert graph by hand into a list of vertices and edges, which you will input into your program
- Write a program to output the
    1. Path Score
    2. The label of the beginning vertex on the path
    3. the label of the ending vertex on the path
    4. (in order) the labels for all the edges that occur on this path.
- Run this program “unconstrained” and with the start and end constraints of “vii” and “i” respectively

Example Input:

```
V vii START
V vi
V v
…
E A ii i -1
E B iii i 5
```

Test Example:



HW4: Program 1

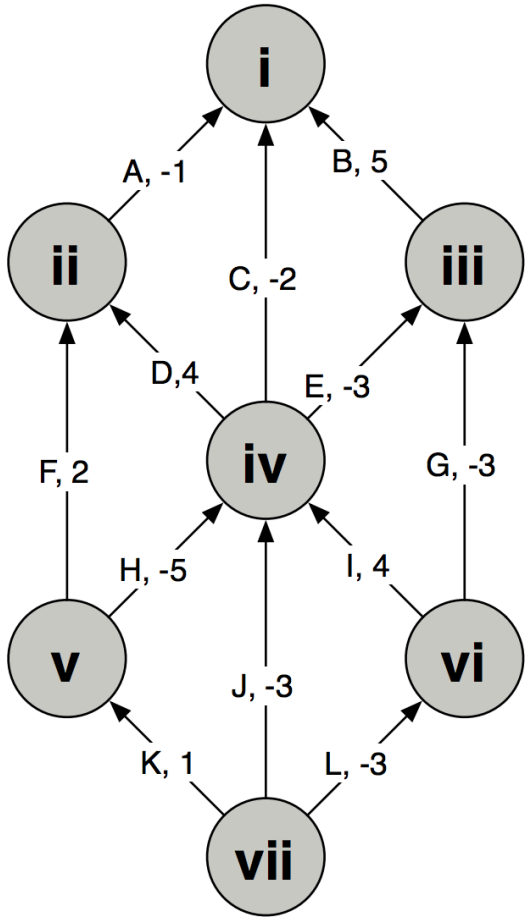
Program 1: Find a highest-weight path in a weighted directed acyclic graph

- Convert graph by hand into a list of vertices and edges, which you will input into your program
- Write a program to output the
    1. Path Score
    2. The label of the beginning vertex on the path
    3. the label of the ending vertex on the path
    4. (in order) the labels for all the edges that occur on this path.
- Run this program “unconstrained” and with the start and end constraints of “vii” and “i” respectively

Example output:

Part 1	Part 2
Score: 8.0	Score: 4.0
Begin: vi	Begin: vii
End: ii	End: i
Path: ID	Path: LIDA

Test Example:



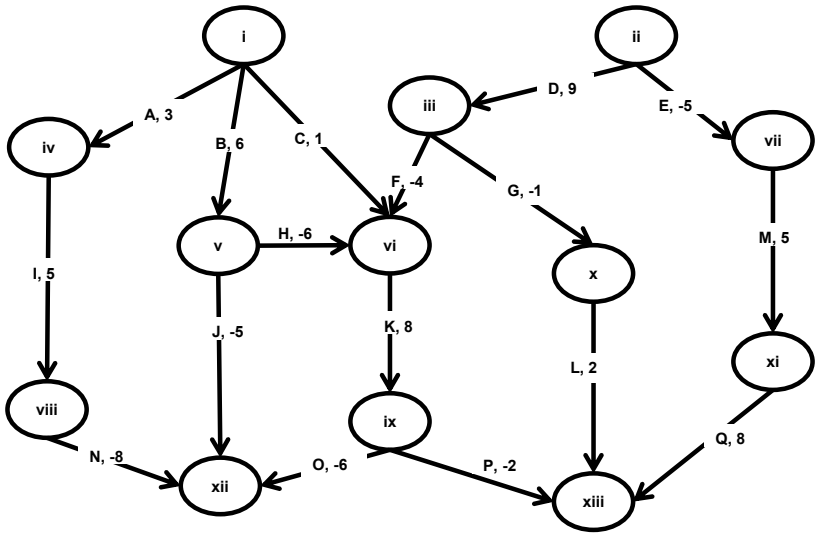
HW4: Program 1

Program 1: Find a highest-weight path in a weighted directed acyclic graph

- Convert graph by hand into a list of vertices and edges, which you will input into your program
- Write a program to output the
    1. Path Score
    2. The label of the beginning vertex on the path
    3. the label of the ending vertex on the path
    4. (in order) the labels for all the edges that occur on this path.
- Run this program "unconstrained" and with the start and end constraints of **"i" and "xiii" respectively**

Output: **Turn this in**

For Grading:



Program 1 Algorithm

- process the v in depth order (*or any order in which parents precede children*)
- if v has no parents, $w(v) = 0$ (the only path ending at v is (v)).
- for any other v, except for the path (v) (which has weight 0), any path ending at v is of form $(v_0, v_1, \ldots, v_k, u, v)$. Then
- u is a parent of v, so $w(u)$ has already been computed, and

$$w((v_0, v_1, \ldots, v_k, u, v)) \leq w(u) + w((u,v))$$

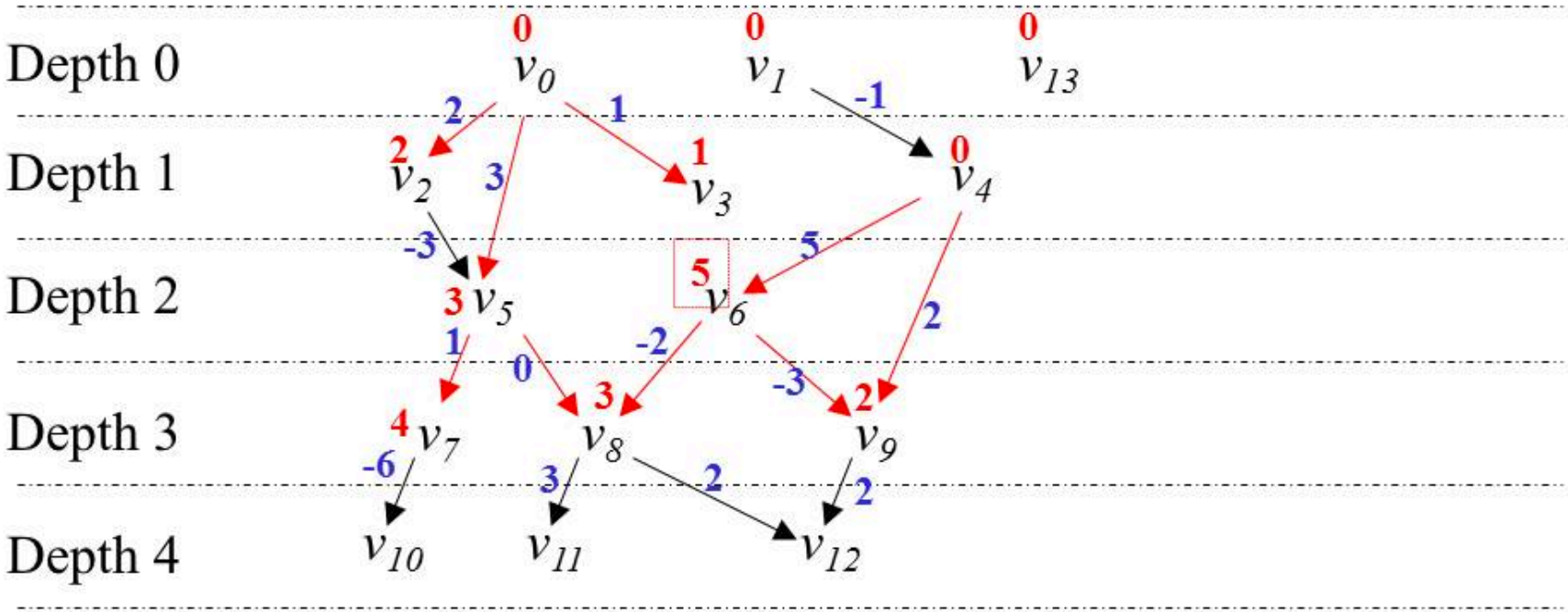
with equality for an appropriate choice of v_i.

- Therefore we may compute $w(v)$ as

$$w(v) = \max(0, \max_{u \in parents(v)} (w(u) + w((u,v))))$$

1) Arranging vertices by depth
2) dynamic programing

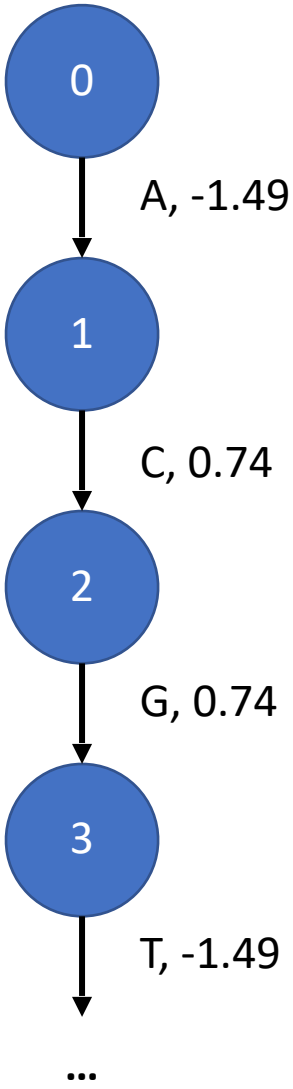
$$w(v) = \max(0, \max_{u \in parents(v)} (w(u) + w((u,v))))$$



HW4: Program 2

Program 2: Convert FASTA to written graph format

- Read 2 input files:
  1) a FASTA file containing a DNA sequence, and
  2) a 'scoring scheme' file that indicates a score to be attached to each possible base (A, C, G, T, or other)
- Output a written graph to run through program 1



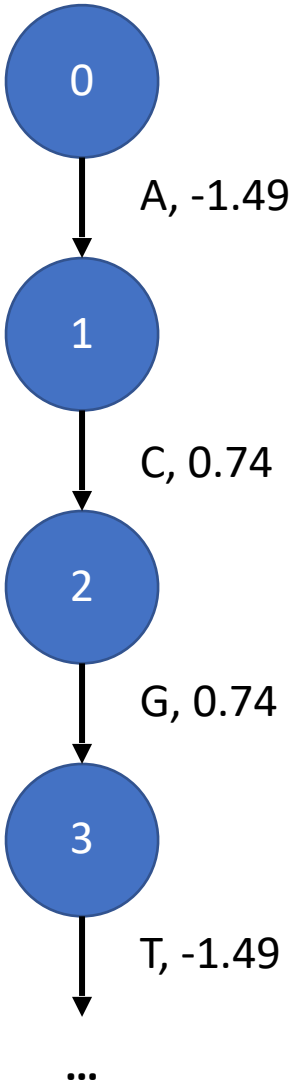
HW4: Program 2

Program 2: Convert FASTA to written graph format

Example of written graph output:

```
V 0
V 1
V 2
…
E A 0 1 -1.49
E C 1 2 0.74
E G 2 3 0.74
```

Run this output through program 1



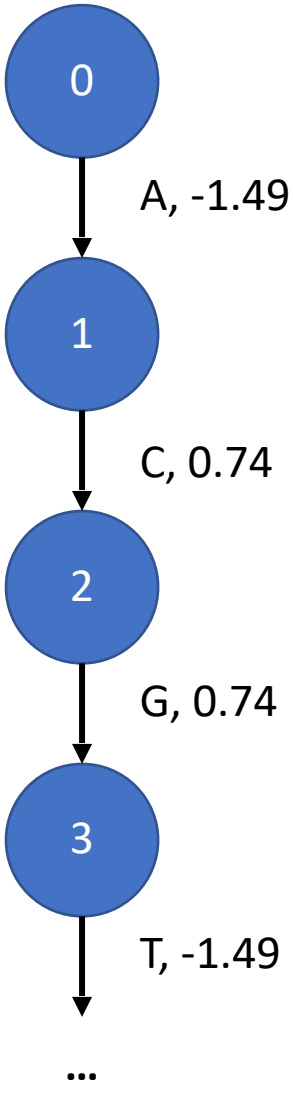
HW4: Program 2

Program 2: Find a highest-weight path in a linked list

Test example: **Mycoplasma gallisepticum**

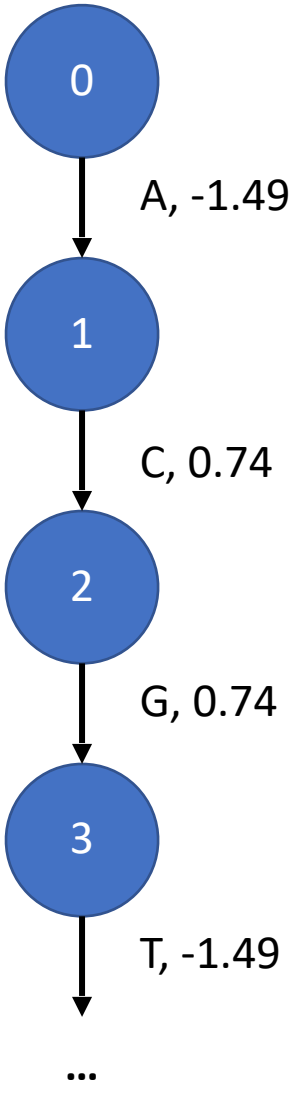
Part 3
>gi|400273702|gb|CP003508.1| Mycoplasma gallisepticum NC96_1596-4-2P, complete genome
*=986257
A=337443
C=156212
T=336693
G=155909
N=0

Score: 11.069998
Begin: 344420
End: 344444
Path: GGCGGCGGCCCCTGGCGATGGCCG
Description: This sequence lies within the HFMG96NCA_2038 gene (encodes a hypothetical protein).



Refer to online hw for more thorough details

HW4: Program 2



Program 2: Find a highest-weight path in a linked list

Test example: **Mycoplasma gallisepticum**

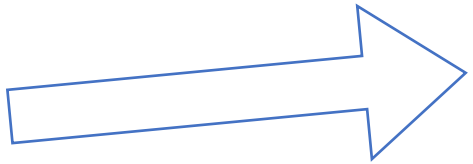
Part 3
>gi|400273702|gb|CP003508.1| Mycoplasma gallisepticum NC96_1596-4-2P, complete genome
*=986257
A=337443
C=156212
T=336693
G=155909
N=0

Score: 11.069998
Begin: 344420
End: 344444
Path: GGCGGCGGCCCCTGGCGATGGCCG
Description: This sequence lies within the HFMG96NCA_2038 gene (encodes a hypothetical protein).

HW2

- Original sequence

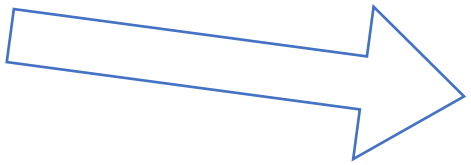
```
Nucleotide Frequencies:
A=0.3053
C=0.1996
G=0.2005
T=0.2947
```



- simulated_markov_0

```
Nucleotide Frequencies:
A=0.3050
C=0.1997
G=0.2005
T=0.2947
```

```
Dinucleotide Frequency Matrix:
A=0.1207 0.0622 0.0587 0.0637
C=0.0650 0.0449 0.0328 0.0569
G=0.0501 0.0455 0.0450 0.0599
T=0.0695 0.0470 0.0640 0.1141
```

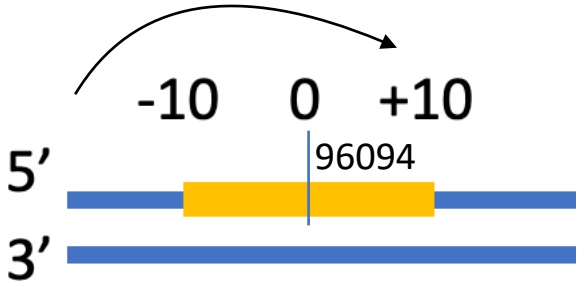


- simulated_markov_1

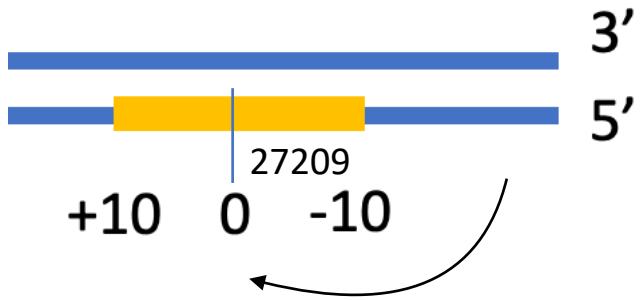
```
Dinucleotide Frequency Matrix:
A=0.1205 0.0620 0.0586 0.0639
C=0.0646 0.0446 0.0330 0.0569
G=0.0502 0.0456 0.0451 0.0600
T=0.0696 0.0469 0.0643 0.1144
```

Any Questions on HW3?

```
CDS             96094..97215
                /locus_tag="PTSG_00022"
                /codon_start=1
                /product="hypothetical protein"
                /protein_id="EGD72006.1"
                /db_xref="GI:326426436"
                /translation="MVVAAGSGGASRPTNAPSCPLCPGGSVGGAVLMVVPLLVCIALL
                AGCLSVSSLWRRNKRQRHAPQYASTCASGRAKPNKRAAPRVQPDLRLPHQQQQPQHPQ..."
```




```
CDS             complement(join(15291..15934,16108..16234,16358..16394,
                16582..16790,17086..17196,17376..17456,17810..17877,
                18020..18060,18199..18256,18556..18598,18767..19187,
                19334..19410,19552..19631,19795..19917,20098..20183,
                20449..20577,20789..20904,21261..21449,21667..21787,
                21936..22108,22453..22549,22808..22934,23895..23970,
                24140..24246,24389..27209))
                /locus_tag="PTSG_11525"
                /codon_start=1
                /product="hypothetical protein"
                /protein_id="EGD71990.1"
                /db_xref="GI:326426420"
                /translation="MWRSWRHGEVGSGVAGGENGKDAQQASSNSHGSHGSHGSNHPNG
                NHGGSSDNVGSSHDERSSSDREQERGQVQRRKRRHARMHEKHASNHAASSVARPSRLT..."
```



Any Questions on HW3?

- Positions downstream of the translation start site could be noncontiguous
  - join(1000…1008, 1200…1500)

−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	0	1	2	3	4	5	6	7	8	9	10
990	991	992	993	994	995	996	997	998	999	1000	1001	1002	1003	1004	1005	1006	1007	1008	1200	1201

- Your results should be the same to the template on the test data

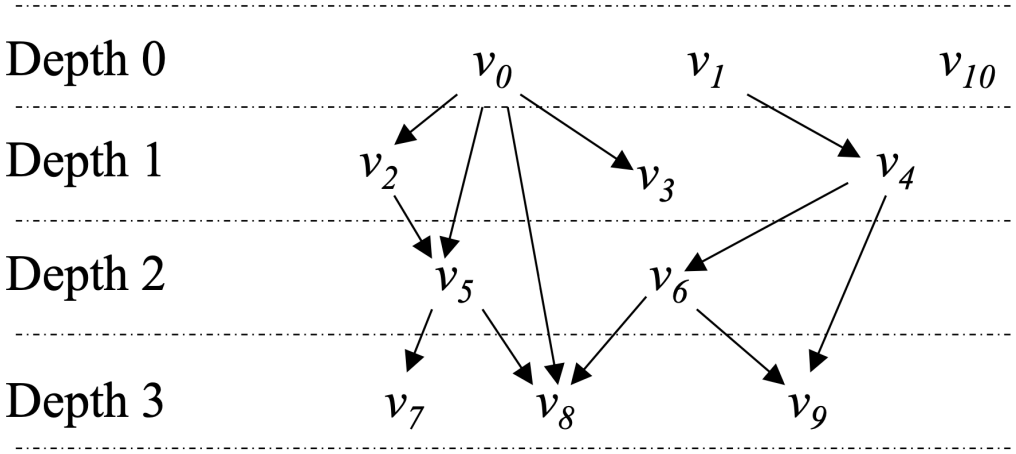
- Check your output format (for example, to 4 decimal places).

The output of HW3 is very big, so the correct format is important.

- We will slightly change the policy on incorrectly formatted assignments - **Incorrectly formatted homework submissions will not be grated. They can be resubmitted, but a late penalty may apply.**

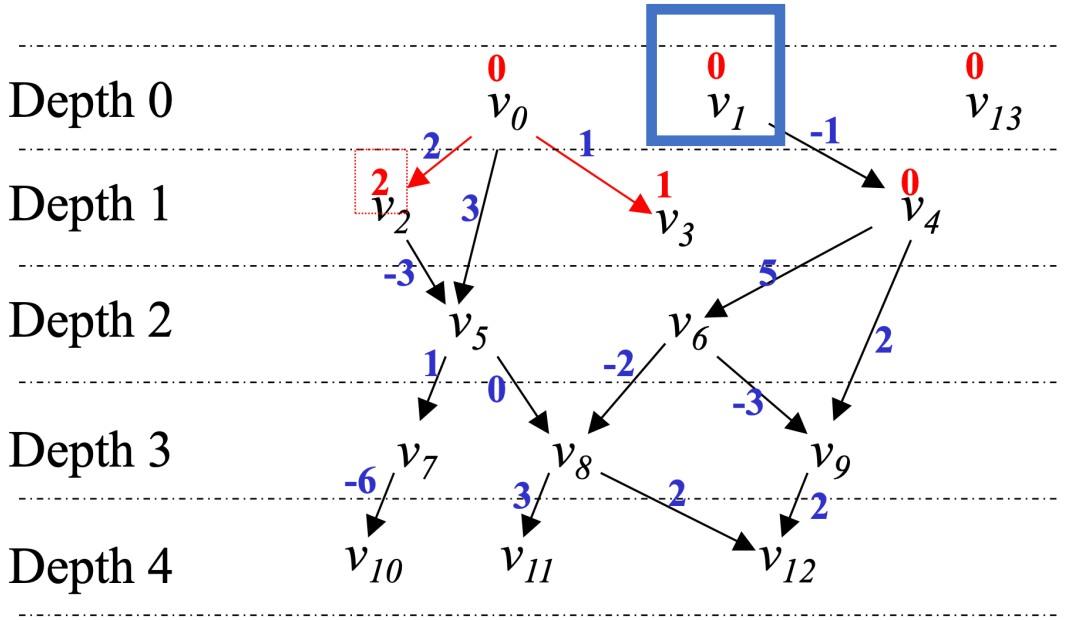
HW4 - Find a highest-weight path in a weighted directed acyclic graph

1) Arranging vertices by depth



2) dynamic programing

$$w(v) = \max(0, \max_{u \in parents(v)} (w(u) + w((u,v))))$$



3) “Constrained”

For example, requiring the path start at node v1