CS 3824
Homework Assignment 6

Given: November 18, 2014 **Due:** December 6, 2014

General directions. The point value of each problem is shown in [ ]. Each solution must include all details and an explanation of why the given solution is correct. In particular, write complete sentences. A correct answer without an explanation is worth no credit. The completed assignment must be turned in as a PDF through Scholar by 5:00 PM on December 6, 2014. **No late homework will be accepted.**

Digital preparation of your solutions is mandatory. Use of LaTeX is optional, but encouraged. No matter how you prepare your homework, **please include your name.**

Use of LaTeX (optional, but encouraged).

- Retrieve this LaTeX source file, named `homework6.tex`, from the course web site.
- Rename the file <*Your VT PID*>`_solvehw6.tex`, For example, for the instructor, the file name would be `heath_solvehw6.tex`.
- Use a **text editor** (such as `vi`, `emacs`, or `pico`) to accomplish the next three steps.
- Uncomment the line

  `% \setboolean{solutions}{True}`

  in the document preamble by deleting the `%`.
- Find the line

  `\renewcommand{\author}{Lenwood S. Heath}`

  and replace the instructor's name with your name.
- Enter your solutions where you find the LaTeX comments

  `% PUT YOUR SOLUTION HERE`
- Convert your solutions to PDF and submit your solutions through Scholar by 5:00 PM on December 6, 2014.

---

[20] 1. Protein Folding. Go to

`http://fold.it/portal/info/science`

register, and download the Foldit Scientific Game for your platform. If you have difficulties installing Foldit and getting it to run, please post to Piazza with your operating system and error message details to get help.

At the Foldit web site, learn what the game is and why solving the protein folding problem is important for mankind. In particular, read these two short articles

`http://www.scientificamerican.com/citizen-science/foldit-protein-exploration-puzzle/`
`http://www.scientificamerican.com/article/foldit-gamers-solve-riddle/`

Once you have Foldit running online, learn about the protein folding process by going through at least 10 intro puzzles. **This may take some time!** These will demonstrate the very basics of structure manipulation tools available in the game. Just follow the prompts. In each case, what you are doing, in effect, is minimizing the energy of the system — the score that is displayed at the top is the "depth" of the current energy minimum; the higher the score the lower (the better) is the energy.

Once you are done with the intro puzzles, click "Science Puzzles" from the main menu, check "show expired" and "show beginner" and then download an actual Science Puzzle: "Easy Mini Freestyle". Use Foldit tools that you have learned to fold this protein — what you want to do is to create a compact, hairpin-looking structure without clashes. It should be fairly easy to get a score of at least 7800. You can stop once this score is achieved, or you can try to get a better score, if you wish. You can save a screenshot from Foldit. Take a screen shot of your highest score solution, and submit it as part of your PDF file. Along with the screenshot, write a half-page discussion (in your own words) of what Foldit is, why protein folding is an important problem, and why it is a hard one.

---

[40] 2. Lloyd's Algorithm. Lloyd's algorithm for k-means clustering always converges to a local minimum of $d(V, X)$. Suppose X^{opt} is a globally optimal set of centers. Then, the *performance ratio* of Lloyd's algorithm is

$$\rho \quad = \quad \frac{d(V, X)}{d(V, X^{\text{opt}})},$$

assuming that $d(V, X^{\text{opt}}) \neq 0$.

You are to show that ρ can be arbitrarily bad (large) for any $k \geq 2$. Start by finding a concrete example for $k = 2$ for which you can make ρ larger than any positive constant. Then, generalize that example to arbitrary $k \geq 2$.

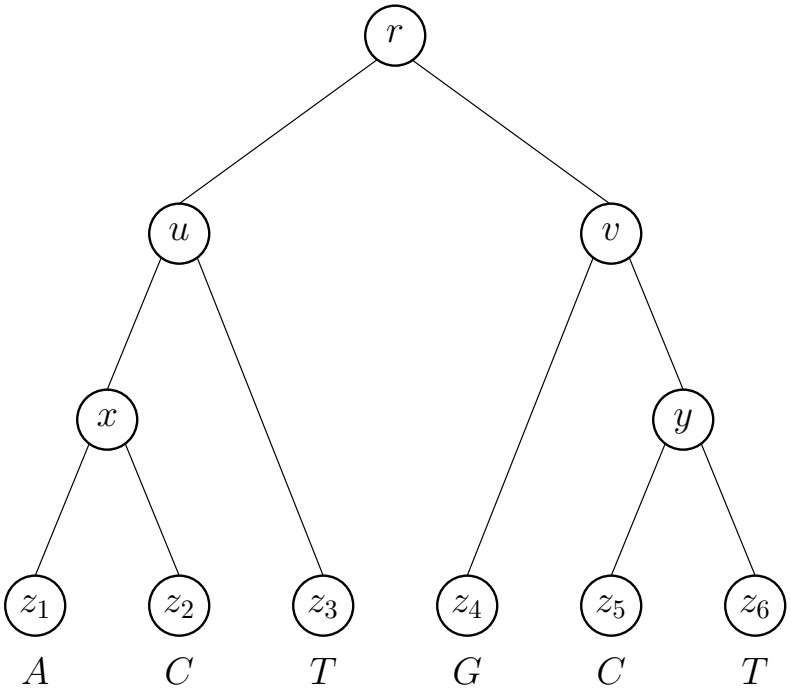

Figure 1: Tree for Exercise 3.

[40] 3. Sankoff's Algorithm. Apply Sankoff's algorithm for the Weighted Small Parsimony problem to the tree in Figure 1. The nodes of the tree are labeled. The observed nucleotides for the leaves are under the leaves. Use this δ function:

δ	A	C	G	T
A	0	3	4	2
C	3	0	1	4
G	4	1	0	3
T	2	4	3	0

Show the calculation of all the s_t values for the leaves and the internal nodes. Compute the actual optimal assignment of nucleotides to internal nodes.