

CSE182 Questions on dictionary matching and regular expressions

Vineet Bafna

October 20, 2008

The questions are open ended, but should help you understand lectures better. Do these questions make sense? Are they helpful in following the lecture? Constructive feedback is appreciated.

1. **Trie construction:** Suppose you were given a query $Q = ACGAAGTCAAGCAGGTA$ as input to BLAST. The “keyword size” parameter was set to 5. Construct an automaton for the keywords. The automaton should have the transition and failure edges. Also, you should describe the lp function.
2. Given the query Q from the previous question, and a random database of size 10^7 nucleotides (all nucleotides independent and equally likely), how many ‘hits’ (exact matches) do you expect to find. What happens if the GC content of the database is 80% ($P[G] = P[C] = 40\%$, $P[A] = P[T] = 20\%$)?
3. Go to NCBI web-site, and run BLAST with a sample query (longer than Q). Search the BLAST output and see if you can locate the number of ‘hits’ reported. It should be at the end of the BLAST search. Does the number of ‘hits’ match your calculations? If not, why not?
4. **Regular expression match** Construct an automaton for the regular expression $R = (D + E)?(S + CC)(C + S)$. A string contains a regular expression R if a substring matches R . Assume that a ? stands for 0 or 1 occurrences of the pattern. Select the strings that contain R among $ASCC$, $DECCC$, and $GCCS$.
5. Extend the automaton for R so as to find strings that contain it. Use this automaton to compute the reachable states $N[i]$ for each string from the previous question, and each position i in the string.
6. Suppose you had a collection of the following words: $AACAA$, $AAGAA$, $AAGAT$, $AAGTA$. To find an exact match to any of these keywords, you can either make a regular expression of these, or a trie. Make both. Which is a better choice? Why?
7. What are the strengths and weaknesses of Profiles versus Regular expressions. The *transmembrane* region of a protein is a stretch of (mostly) hydrophobic residues that pass through the membrane. If you are modeling transmembrane regions, are regular expressions better, or profiles? How about HMMs (assuming you have studied L7)?
8. Describe ψ -Blast. (Look up the NCBI web-site). Explain why ψ -blast could be more sensitive than regular Blast.
9. Search the NCBI page for ϕ -Blast, a tool that we did not cover in class. Try to understand how ϕ -Blast works. When would you use ϕ -Blast? Also, what do the tools TBlastn, Blastx, and Tblastx do?