Algorithms in Bioinformatics
Fall 2014

Hands on practices

Week 2: Run a typical Bioinformatics search program

1. WU-BLAST

1. 1 What is WU-BLAST?

A once popular version of BLAST.

It can be accessed in /share/home/ccwei/tools/wu-blast

1.2 Run WU-BLAST

The query sequence is the human U1A RNA binding protein, and its sequence is here (/share/home/ccwei/pab/2014/week2/u1_human.fa.gz). Get the sequence and save it to a file called u1_human.fa. The database is C.elegans 215, the complete proteome of C.elegans (*/share/home/ccwei/pab/2014/week2/C.elegans/Proteome/*ws_215.protein).

1.2.1 Go to a working directory. For example, ~/week1. Note, you should choose a directory for your own.

1.2.1 Create a protein sequence database using xdformat program.

export BLAST=/share/home/ccwei/tools/wu-blast

export DATA=/share/home/ccwei/pab/2014/week2/C.elegans/Proteome/

export PATH=$PATH:$BLAST

$BLAST/xdformat -p -o worm_protein $DATA/ws_215.protein.fa

1.2.2 Check the database files

ls

There will be three files, ending with .xpd,.xps and xpt.

1.2.3 Run BLAST search.

```
$BLAST/blastp                                worm_protein
/share/home/ccwei/pab/2014/week2/u1_human.fa  filter=seg+xnu > blast.out
```

2. Parsing Smith/Waterman output (Perl)

You also have a legacy Perl script that takes a WU-BLAST output file and parses it to find the name of the query seuqence, and the name, socre and P-value of the top scoring hit. The source code for the script is here (/share/home/ccwei/pab/2014/week2/blastparser.pl) . An example of WU-BLAST output is here (/share/home/ccwei/pab/2014/week2/blast.out) . Make a copy of this script and the output in files called blastparser.pl and blast.out, and make the parser executable as a program (chmod +x blastparser.pl). When you run the script on the sample output file, it produces a single summary line of output as follows:

/share/home/ccwei/pab/2014/week2/blastparser.pl blast.out

```
Best hit to RU1A_HUMAN is: K08D10.3, with score 378, P-value
3.2e-53.
```