SE498 Parallel Computing
Lab 4: Using MPI to aid in the sequencing of genes
Due: October 30, 2013

1. Objectives

- Construct an MPI program which aids in the sequencing of DNA
- Perform rudimentary analysis of biomolecular proteins
- Analyze the performance improvement of parallelizing a system

2. Introduction

DNA represents the essential science behind life. DNA describes the relationships between a homosapien, a Microtus pennsylvanicus (common field mouse), and a Loxodonta Africana (African Elephant). Scientists studying in the biological field sue massive software packages to aid in the analysis of DNA sequencing.

One of the largest scientific projects ever undertaken is the human genome project. Quoting from http://www.genome.gov/12011238

> *"The Human Genome Project (HGP) was the international, collaborative research program whose goal was the complete mapping and understanding of all the genes of human beings. All our genes together are known as our `genome.'*
>
> *The HGP was the natural culmination of the history of genetics research. In 1911, Alfred Sturtevant, then an undergraduate researcher in the laboratory of Thomas Hunt Morgan, realized that he could - and had to, in order to manage his data - map the locations of the fruit fly (Drosophila melanogaster) genes whose mutations the Morgan laboratory was tracking over generations. Sturtevant's very first gene map can be likened to the Wright brothers' first flight at Kitty Hawk. In turn, the Human Genome Project can be compared to the Apollo program bringing humanity to the moon.*
>
> *The hereditary material of all multi-cellular organisms is the famous double helix of deoxyribonucleic acid (DNA), which contains all of our genes. DNA, in turn, is made up of four chemical bases, pairs of which form the "rungs" of the twisted, ladder-shaped DNA molecules. All genes are made up of stretches of these four bases, arranged in different ways and in different lengths. HGP researchers have deciphered the human genome in three major ways: determining the order, or "sequence," of all the bases in our genome's DNA; making maps that show the locations of genes for major sections of all our chromosomes; and producing what are called linkage maps, complex versions of the type originated in early Drosophila research, through which inherited traits (such as those for genetic disease) can be tracked over generations.*
>
> *The HGP has revealed that there are probably about 20,500 human genes. The completed human sequence can now identify their locations. This ultimate product of the HGP has given the world a resource of detailed information about the structure, organization and function of the complete set of human genes. This information can be thought of as the basic set of inheritable "instructions" for the development and function of a human being.*
>
> *The International Human Genome Sequencing Consortium published the first draft of the human genome in the journal Nature in February 2001 with the sequence of the entire genome's three billion base pairs some 90 percent complete. A startling finding of this first draft was that the number of human genes appeared to be significantly fewer than previous estimates, which ranged from 50,000 genes to as many as 140,000.The full sequence was completed and published in April 2003."*

DNA sequencing is a laboratory technique used to determine the exact sequence of the four base chemicals present in DNA. These four base chemicals are adenine (A), cytosine (C), guanine (G), and thymine (T). The DNA base sequence carries the information a cell needs to assemble protein and RNA molecules.

Like many computer datasets, there are multiple formats for the storage of DNA data. For our purposes, we are going to use the FASTA format. In the FASTA format, a sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line must begin with a greater-than (">") symbol in the first column. Lines that follow contain 60 or less DNA characters.

An example sequence in FASTA format is:

```
>AB000263 |acc=AB000263|descr=Homo sapiens mRNA for prepro cortistatin like peptide,
complete cds.|len=368
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCG
CTGCCCTGCCCCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGG
CAGGAATAAGGAAAAGCAGCCTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCC
AGGCCAGTGCCGGGCCCCTCATAGGAGAGGAAGCTCGGGAGGTGGCCAGGCGGCAGGAAG
GCGCACCCCCCCAGCAATCCGCGCGCCGGGACAGAATGCCCTGCAGGAACTTCTTCTGGA
AGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTCACGCAAGTTTAATTACA
GACCTGAA
```

3. Lab activities

For this lab, you are going to write a program for the Glenn supercomputer which will aid in analyzing DNA sequences.

The program has two distinct uses. The first case involves counting the number of sequences of a given length present within the DNA file. For example, lets use the following DNA sequence:

```
>Test
ACAAGATG
```

If we would like to know how many times each protein appears in the sequence, we can use a sequence length of 1. Running the program should give the following results:

```
./analyzeSequence test.dna -L1
Looking for sequences of length 1
A -  4
C - 1
T - 1
G - 2

./analyzeSequence test.dna -L2
Looking for sequences of length 2
AC - 1
CA - 1
AA - 1
```

```
AG - 1
GA - 1
AT - 1
TG - 1
```

Overall, we would like to know the speedup as we search for different sequence lengths in a given file and vary the number of processors running. For example, maybe we desire to know how many different sequences of length 10 exist in the overall DNA sequence. How long will this take to do?

The second thing we desire to do is to be able to search for a given sequence in the DNA strand. Assuming the test sequence again, we might execute the following:

```
./analyzeSequence test.dna –Stestseq.dna
The sequence AAGA (Some random sequence described in a FASTA
file) occurs once in the given test file starting at location 2.
```

4. Basic Design

The specifics of the design and implementation of this program are up to you. However, the general design should use an MPI scatter to split the DNA sequence across n processing nodes and then some form of either a gather or other operation to combine the results back together. The algorithm should be efficient, and should achieve as high of a speedup as is possible.

5. Basic Steps

The course website has two long DNA samples available for testing. The first sample represents a modified human DNA sequence. The second is a shorter DNA sequence which is made up for testing purposes. The testing file allows you to achieve results in a quicker fashion, while the longer file is intended for your actual analysis.

The first part is to determine the counts of DNA. For the long sequence, you should determine how often each sequence of length 1 through 4 occurs, as well as what speedup and efficiency occur if the problem is executed on 1 through 8 processing cores.[1]

For the second part, there are a set of DNA sequences that are to be searched for in the given DNA file. A search might be useful, for example, to detect if the BRCA gene is present in a person's DNA. (BRCA is linked to an increase in Breast and ovarian cancers in women.) There are a few short samples that you are to search for of varying length. Select a few different sizes and vary their usage from 1 to 8 processors, plotting the speedup and efficiency as the code changes.

[1] Hint: It probably is wise to write all of your simulations into a script and submit a single script, thus gathering all of your data at once. However, make certain that the tool is working properly before doing this, as it could yield unintended consequences.

6. Deliverables

Each lab team is responsible for submitting a report with the following

1. What did you learn by doing this lab?
2. What things went right and wrong when doing this lab?
3. For the sequence length analysis,
    a. How many sequences of length 1 through 4 exist in the code, and how often do they occur
    b. What was the execution time for each analysis as you ran on between 1 and 8 processors
    c. What was the speedup and efficiency as the number of processors changed.
4. For the search,
    a. For each of the substrings you choose to search for,
        i. Was the substring found in the DNA?
        ii. How long did it take to find the substring with between 1 and 8 processors running
        iii. What was the speedup?
        iv. What was the efficiency?
5. Explanation: Explain how you handled sequences and searches that crossed the boundaries between processors. For example, if you used a gather on the sequence "AAATGTGA" and tried to have 8 processors, each processor would receive 1 protein. How would you handle looking for the gene sequence TG then?
6. What conclusions can you draw from this experience?

Additionally, each team is responsible for submitting a tar.gz file which contains all code and executable and a working job script which can be submitted on the OSC cluster.

Reference Links:
http://useast.ensembl.org/info/data/ftp/index.html
http://useast.ensembl.org/info/data/ftp/index.html
http://www.ncbi.nlm.nih.gov/nuccore/DI053969.1
http://www.genomatix.de/online_help/help/sequence_formats.html
http://www.ncbi.nlm.nih.gov/guide/data-software/#downloads