Projeto de Bioinformática

Neste projeto, os alunos desenvolverão um sistema para análise de sequências genéticas de DNA utilizando técnicas de programação paralela.

A análise de DNA é crucial no mundo real para identificar mutações genéticas, estudar doenças, desenvolver tratamentos e melhorar a compreensão da biologia humana e de outras espécies.

Sequências genéticas podem ter bilhões de bases, como o genoma humano, que contém cerca de 3 bilhões de pares de bases. Processar tais volumes de dados de forma eficiente exige paralelismo para reduzir significativamente o tempo de processamento e tornar a análise viável.

Sobre o contexto de trabalho

Bases nucleotídicas são os blocos construtivos do DNA e RNA, que formam as sequências genéticas responsáveis pelo armazenamento e transmissão da informação genética em todos os organismos vivos.

As bases nucleotídicas no DNA são compostas por quatro tipos: **Adenina (A), Timina (T), Citosina (C)** e **Guanina (G).** Essas bases se pareiam de forma específica:

- A sempre se liga com T, e
- C sempre se liga com G,

formando a estrutura de dupla hélice característica do DNA. Essas sequências codificam informações essenciais para a síntese de proteínas e funções celulares.

No RNA as bases nucleotídicas têm uma diferença em relação ao DNA. Enquanto o DNA contém Adenina (A), Timina (T), Citosina (C) e Guanina (G), **o RNA possui Uracila (U) em vez de Timina (T)**. As bases no RNA, portanto, são Adenina (A), Uracila (U), Citosina (C) e Guanina (G).

Sobre os dados

Trabalharemos com arquivos FASTA que contêm uma versão modificada da montagem do genoma humano de referência de fev. 2009 (GRCh37/hg19). As sequências cromossômicas foram montadas pelos centros de sequenciamento do Projeto Genoma Humano Internacional.

O que é um arquivo FASTA?

Um arquivo FASTA é um formato de arquivo amplamente utilizado em bioinformática para representar sequências de nucleotídeos (DNA/RNA) ou aminoácidos (proteínas). Cada

entrada em um arquivo FASTA começa com uma linha de descrição, precedida por um sinal de maior (">"), que contém informações sobre a sequência, como o identificador e detalhes adicionais. As linhas subsequentes contêm a sequência biológica em si, dividida em blocos para facilitar a leitura. Esse formato é popular devido à sua simplicidade e compatibilidade com diversos softwares de análise genética.

Como obter os dados?

Para baixar os dados, faça:

```
wget "ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/snp147Mask/chrX.subst.fa.gz"
```

onde: X varia de 1 até 22.

Para descompactar os arquivos, faça:

```
gunzip chrX.subst.fa.gz
```

onde: X varia de 1 até 22.

DICAS:

- programe e debugue utilizando dois a quatro arquivos. Depois deixe rodar com todos os arquivos!
- Conversa as bases todas para letras minúsculas ou maiúsculas antes de realizar as análises!

```
258516  TGTAAGGCAGCAGCCCGAGCTCTGCTCATGGCTGGTCTYCAGCAGCTGCC
258517  AGCGGGCTGGCCCTCCTCCCTCATCAGCCATGCTGAGCATTCCCTGCCTT
258518  TTCGGGAGAACACGMGGTCASGCKGGCTGCAGTAAGGCAGGRGGAAGGAG
258519  CAGGCGGAAATGCTGAGGAACTCCTTGGTTCTCTCACCCCAGAACTCTTC
258520  CCAGCCTGCYAGGCTCTgggcctgagggatcattcacccaccaatcccac
258521  cccgtgacttagtcatggaggaagtgaggyctggagagaggggacttgtc
258522  ccagCAGGGATTAGGGCTTTCTTCCCTGCTGTGTGGCAGAYGCTCCTGCA
258523  CCGAATTCTGCSATGCYACGCACCACATGCTCACTGGCRCCTTTTTYTGGG
258524  TTCTCAGTGCTGCCCCCACAAGYTTAACTTGTAGCTGGGGAARACCAGCA
258525  GAGGTCRGAACAATTRATACAATYACGYAGCCTGAGAGASAGCARYAGTG
258526  CTAYGSWAAAARAGTACTTCaattwaaaaattctaattgtaaaaataaat
258527  tttaaaGAGAACTGGCCCAGGACATCCATGCCACATTYCCTCTTCCTCCA
258528  GGCCCTTCCTGACGCCCTTGACTCCGGCCTGTCCTGAGGCATGCAGATGC
258529  CACTCACCTCCAAGTTCTCCAGAACCTCTGGGACACCAAAGACCCTGAYC
258530  TCGGAGGTGGTGTAGTGGTTGCTCRGAAGGATTTCAGACTGCTTGGCATA
258531  GAGATCTGGGCTGAAGGSYACTGgccagctggcaaatgacacggagttca
258532  gccttgtsctcacacagcgagggaaggacggggaggtttggagctgayag
258533  tcaacaaatgggtaatgggcGGCCATTAGTGAAGTGGCAGGAAGTGGGGC
258534  ASCACACACAGYGCGAGTTGTTTTGTGTATGAACRGCCAGGCGACGRTGT
258535  CTGAACATCTTTCCTTAATTTTGCGAACGGAARCAGGGCACGATGAACCT
258536  GGGAACTCACAAAAATGATGAGCATAGAGCTCTGCGGGGTGCGGTAAGAA
```

Figura 1. Exemplo de parte de um arquivo FASTA

Mais informações no site:
https://hgdownload.cse.ucsc.edu/goldenPath/hg19/snp147Mask/

Sobre as tarefas

Siga as atividades na ordem, pois você verá que uma viabiliza as demais! Você pode explorar os recursos computacionais e outras estratégias que julgar pertinentes, como salvar arquivos intermediários pelos nós MPI. Qual o comportamento disso? Todas as máquinas ficam sincronizadas?

Exercício 1: Contagem de Bases

Descrição: Implemente um programa paralelo que conte o número de ocorrências de cada base (A, T, C, G) em uma grande cadeia de DNA.

- MPI: Divida a cadeia entre processos diferentes e agregue os resultados ao final. Como dividir? Parte dos dados? Uma porção dos arquivos por máquina?
- OpenMP: Use paralelização em laços para distribuir a contagem entre threads. Como consolidar a contagem? Um dicionário? Um array? Variáveis soltas?

Exercício 2: Transcrição de DNA em RNA

Descrição: Desenvolva um programa que converta uma sequência de DNA em RNA. Lembre das substituições das bases nitrogenadas!

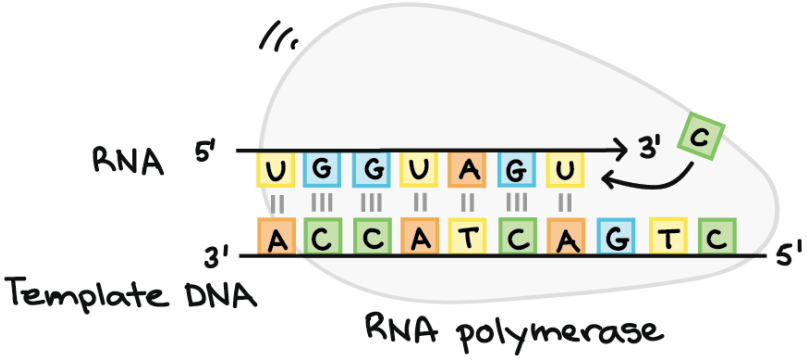

Figura 2. Relações entre pares de bases nitrogenadas entre DNA e RNA

- MPI: Distribua a conversão entre diferentes processos e una as partes convertidas. Será que não vale a pena salvar o RNA em arquivos? Há parte dos dados que podemos ignorar?
- OpenMP: Paralelize a substituição em um loop, dividindo as tarefas entre threads.

Exercício 3: Trabalhando com aminoácidos

O mRNA (RNA mensageiro) serve como um intermediário que transporta a informação genética do DNA no núcleo para os ribossomos no citoplasma, onde ocorre a tradução em proteínas.

Um códon é uma sequência de três nucleotídeos em DNA ou RNA que especifica um aminoácido ou um sinal de parada durante a síntese de proteínas. No processo de tradução, **o códon de início é AUG**, que codifica a metionina e sinaliza o início da tradução da proteína. **Os códons de término típicos são UAA, UAG** e **UGA**, que não codificam aminoácidos e sinalizam o fim da síntese proteica.

Descrição: A partir da conversão de DNA → RNA, faça um programa que conte quantas proteínas foram inicializadas (contagem de AUG). Faça a distribuição paralela que julgar pertinente.

Exercício 4: Trabalhando com síntese proteica

Descrição: Desenvolva um programa que percorra uma sequência de RNA e traduza cada códon em seu aminoácido correspondente, até encontrar um códon de parada.

Códon (RNAm)	Aminoácido
CCA, CCG, CCU, CCC	Prolina
UCU, UCA, UCG, UCC	Serina
CAG, CAA	Glutamina
ACA, ACC, ACU, ACG	Treonina
AUG	Metionina (início)
UGA	Códon STOP
UGC, UGU	Cisteína
GUG, GUA, GUC, GUU	Valina

Figura 3. Identificando aminoácidos

DICA: represente cada aminoácido por um número, e gere uma sequência de números para representar a tradução.

- MPI: Divida a sequência de RNA em partes entre processos, cada um traduzindo uma parte e enviando o resultado ao processo mestre para montar a proteína completa. O que faz mais sentido? Separar por arquivos ou parte dos arquivos?
- OpenMP: Divida a tarefa entre threads para processar a tradução de forma simultânea. O que faz mais sentido? Testar cada busca por aminoácido numa thread?

Entrega

- Submeta seus **códigos "cpp"** e **".slurm"** no BB! A entrega ocorrerá até 15/nov.