# BCB410 HMM Exercises

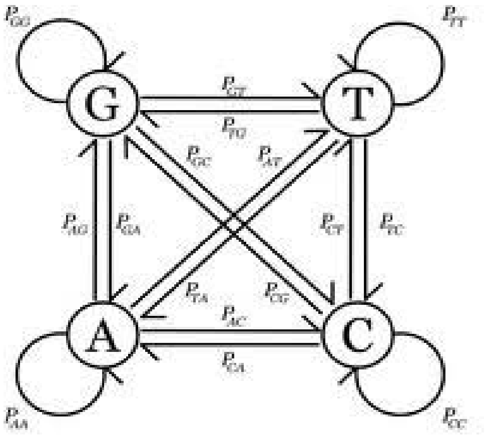
- **Statistics behind the (Hidden) Markov Model:**

**The following two questions prove that the (Hidden) Markov Model really describes a proper probability distribution over the whole space of sequences.**

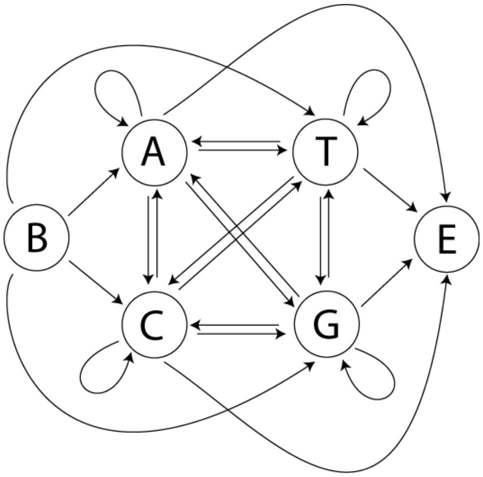
1. **Based on the below markov model, prove the sum of the probability of all possible sequences of length L is equal to 1.**

   **Remeber the probability for sequence X with length L is**

   $\mathbf{P(X) = P(X_L|X_{L-1}, \ldots \ldots X_2|X_1, X_1) = \ P(X_L|X_{L-1})P(X_{L-1}|X_{L-2}) \ldots \ldots P(X_1)}$



2. **Assume that the model has an end state (as below), and that the transition from any state to the end state has probability ε. Show that the sum of the probability over all sequences of length L (and properly terminating by making a transition to the end state) is ε(1- ε)L-1. Use this result to show that the sum of the probability over all possible sequences of any length is 1. Hint: Use the result that, for $\mathbf{0 < x < 1, \sum_{i=0}^{\infty} x^i = 1/(1-x)}$**



- **Hmmer3 deployment and use:**

1. **Download Hmmer3 package from http://hmmer.janelia.org/ and install on your own computer. Also check the user guide ftp://selab.janelia.org/pub/software/hmmer3/3.0/Userguide.pdf , you may need it for the following deployment and use.**
2. **Download the newest Pfam database from ftp://ftp.sanger.ac.uk/pub/databases/Pfam/releases/Pfam25.0/Pfam-A.full.gz.**
3. **Download phage database from http://dl.dropbox.com/u/41884121/Joe/phagedb**
4. **Use hmmpress command to index and press your Pfam database to binary file.**
5. **Use hmmfetch command to retrieve the following profile hmms**
   - **Phage_Nu1(PF07471.6) :** Phage DNA packaging protein Nu1
   - **Terminase_2(PF03592.10):** Terminase small subunit
   - **Terminase_4(PF05119.6):** Phage terminase,small subunit
6. **Use hmmstat command to display summary statistics for each profile hmm.**
7. **Use each profile hmm search against phage protein database using hmmsearch command. Check the output file, especially the E-value and score column. Record the number of hits for each profile hmm (E-value cutoff 10.0, you can use different E-value cutoffs). Also record the sum of hits from all three hmm searches.**
8. **Download the seed sequences (no gaps) of each profile hmm from Pfam website and align them all together (using whatever MSA tools you like). Then use Hmmbuild command to build your own hmm bcb410.hmm.**
9. **Use bcb410.hmm search against the phage database and check the output file. Count the number of hits on the output and compare to number you get from step 7 (the sum of hits from all three hmm searches). Observe and explain the differences between these two numbers. Which one is larger and why?**

# BCB410 HMM Exercises

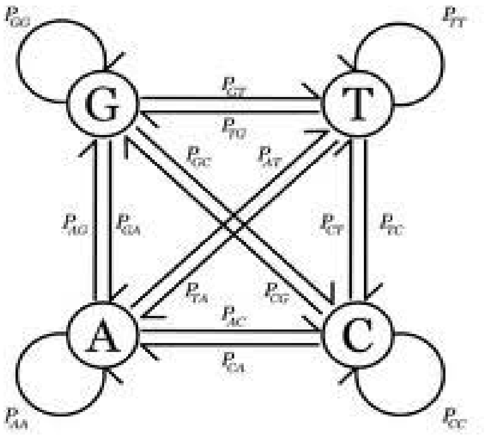
- **Statistics behind the (Hidden) Markov Model:**

**The following two questions prove that the (Hidden) Markov Model really describes a proper probability distribution over the whole space of sequences.**

1. **Based on the below markov model, prove the sum of the probability of all possible sequences of length L is equal to 1.**

**Remeber the probability for sequence X with length L is**

$$\mathbf{P(X) = P(X_L|X_{L-1}, \dots\dots X_2|X_1, X_1) = \ P(X_L|X_{L-1})P(X_{L-1}|X_{L-2}) \dots\dots P(X_1)}$$



**Answer:**

For a particular $X_1$, the sum of transition probability for all possible $X_2$ is 1.

$$\sum_{\{X\}} P(X) = \sum_{X_1}\sum_{X_2} \dots \sum_{X_L} P(X_L|X_{L-1})P(X_{L-1}|X_{L-2}) \dots\dots P(X_1)$$

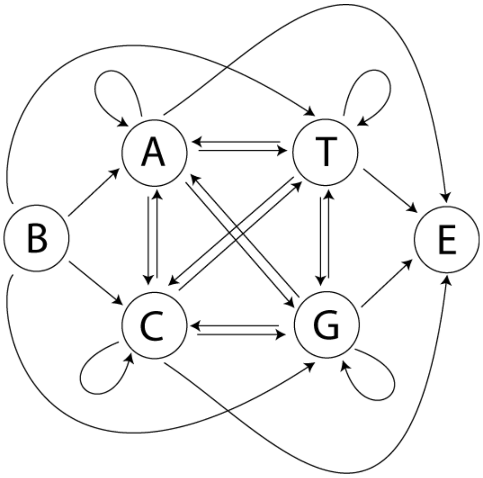
**Note that**

$$\sum_{X_L} P(X_L|X_{L-1}) = 1 \ , \sum_{X_{L-1}} P(X_{L-1}|X_{L-2}) = 1 \ , \qquad \dots \qquad \sum_{X_2} P(X_2|X_1) = 1$$

**Therefore**

$$\sum_{\{X\}} P(X) = \sum_{X_1}\sum_{X_2} \dots \sum_{X_L} P(X_L|X_{L-1})P(X_{L-1}|X_{L-2}) \dots\dots P(X_1) = \sum_{X_1}\sum_{X_2} P(X_2|X_1)P(X_1) = \sum_{X_1} P(X_1) = 1$$

**2. Assume that the model has an end state (as below), and that the transition from any state to the end state has probability ε. Show that the sum of the probability over all sequences of length L (and properly terminating by making a transition to the end state) is $\varepsilon(1-\varepsilon)^{L-1}$. Use this result to show that the sum of the probability over all possible sequences of any length is 1.** **Hint: Use the result that, for $0 < x < 1, \sum_{i=0}^{\infty} x^i = 1/(1-x)$**



**Answer:**

- **For length L (L>0)**

$$\sum_{\{X,L\}} P(X) = \sum_{X_1}\sum_{X_2}\dots\sum_{X_L} P(E|X_L)P(X_L|X_{L-1})P(X_{L-1}|X_{L-2})\dots\dots P(X_1)$$

**Note that**

$$\sum_{X_1} P(X_1) = 1 \textbf{ and } P(E|X_L) = \varepsilon$$

**The sequence with length L start from $X_1$ and end at $X_L$, it has at least one residue.**

$$\sum_{X_L} P(X_L|X_{L-1}) = 1-\varepsilon\ ,\dots \quad \sum_{X_2} P(X_2|X_1) = 1-\varepsilon$$

**For a particular $X_1$, the sum of transition probability for all possible $X_2$ (except the end state E) is 1- ε.**

**Therefore**

$$\sum_{\{X,L\}} P(X) =$$

$$\sum_{X_1}\sum_{X_2}\dots\sum_{X_L} P(E|X_L)P(X_L|X_{L-1})P(X_{L-1}|X_{L-2})\dots\dots P(X_1) = \varepsilon(1-\varepsilon)^{L-1}$$

- **For any length**

$$\sum_{\{X\}} P(X) = \sum_{L=1}^{\infty}\sum_{\{X,L\}} P(X) = \sum_{L=1}^{\infty}\varepsilon(1-\varepsilon)^{L-1} = \sum_{L-1=0}^{\infty}\varepsilon(1-\varepsilon)^{L-1} = \varepsilon\times\frac{1}{\varepsilon} = 1$$

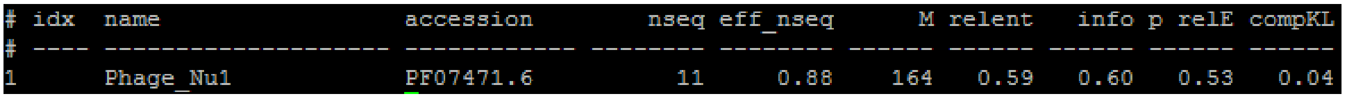
- **Hmmer3 deployment and use:**

1. **Download Hmmer3 package from http://hmmer.janelia.org/ and install on your own computer. Also check the user guide ftp://selab.janelia.org/pub/software/hmmer3/3.0/Userguide.pdf , you may need it for the following deployment and use.**
2. **Download the newest Pfam database from ftp://ftp.sanger.ac.uk/pub/databases/Pfam/releases/Pfam25.0/Pfam-A.full.gz.**
3. **Download phage database from http://dl.dropbox.com/u/41884121/Joe/phagedb**
4. **Use hmmpress command to index and press your Pfam database to binary file.**
5. **Use hmmfetch command to retrieve the following profile hmms**
    - **Phage_Nu1(PF07471.6) :** Phage DNA packaging protein Nu1
    - **Terminase_2(PF03592.10):** Terminase small subunit
    - **Terminase_4(PF05119.6):** Phage terminase,small subunit
6. **Use hmmstat command to display summary statistics for each profile hmm.**
7. **Use each profile hmm search against phage protein database using hmmsearch command. Check the output file, especially the E-value and score column. Record the number of hits for each profile hmm (E-value cutoff 10.0, you can use different E-value cutoffs). Also record the sum of hits from all three hmm searches.**
8. **Download the seed sequences (no gaps) of each profile hmm from Pfam website and align them all together (using whatever MSA tools you like). Then use Hmmbuild command to build your own hmm bcb410.hmm.**
9. **Use bcb410.hmm search against the phage database and check the output file. Count the number of hits on the output and compare to number you get from step 7 (the sum of hits from all three hmm searches). Observe and explain the differences between these two numbers. Which one is larger and why?**

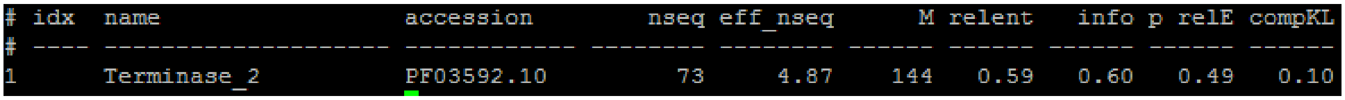
**Answer:**

- **Step1-5: Just follow the instruction.**
- **Step 6: Using hmmstat to display the statistics for each hmm.**

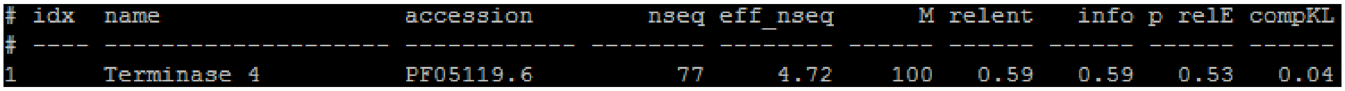
You will find the result looks like the following:



```
# idx  name                 accession        nseq eff_nseq      M relent   info p relE compKL
# ---- -------------------- ------------ -------- -------- ------ ------ ------ ------ ------
1      Phage_Nu1            PF07471.6          11     0.88    164   0.59   0.60   0.53   0.04
```



```
# idx  name                 accession        nseq eff_nseq      M relent   info p relE compKL
# ---- -------------------- ------------ -------- -------- ------ ------ ------ ------ ------
1      Terminase_2          PF03592.10         73     4.87    144   0.59   0.60   0.49   0.10
```



```
# idx  name                 accession        nseq eff_nseq      M relent   info p relE compKL
# ---- -------------------- ------------ -------- -------- ------ ------ ------ ------ ------
1      Terminase_4          PF05119.6          77     4.72    100   0.59   0.59   0.53   0.04
```

**Note:** the nseq indicates the number of seed sequences for each hmm.

- **Step 7: Use each profile hmm search against phage protein database using hmmsearch command.**

You can find the hit result for each hmm on the following files attached in the email (E-value cutoff 10)

1. Phage_Nu1.out (**53 hits**)
2. Terminase_2.out (**76 hits**)
3. Terminase_4.out (**73 hits**)

In total there are **202 hits** for all three hmms.

- **Step 8: use hmmbuild command to build your own hmm bcb410.hmm.**

First download seed sequences (**in total 161**) for all three hmms from pfam, then align them in jalview using Mafft, you may use your desired MSA method. The alignment was saved as fasta format then converted into Stockholm format (**bcb410.sto**) which can be recognized by **hmmbuild.** Use bcb410.sto as an input to Hmmbuild you will get your own hmm as **bcb410.hmm**.

http://sequenceconversion.bugaco.com/converter/biology/sequences/fasta_to_stockholm.php

is the place to convert your fasta to Stockholm. **bcb410.sto** and **bcb410.hmm** are also attached in the email.

- **Step 9: Use bcb410.hmm search against the phage database and check the output file.**

The hit result for **bcb410.hmm** against phage protein database was saved in file **bcb410.out.** This time there are **113 hits** which is significantly lower than **202 hits** we got from step7. The reason is that bcb410.hmm is more specific than the union of three hmms. You may think the bcb410.hmm as the intersection of these three hmms. Although all three hmms are categorized as small terminase for phage, but they don't have lots of sequence similarity, therefore simply combine their seed sequences and make a combined hmm won't improve homology searching, in fact, it will reduce the performance.