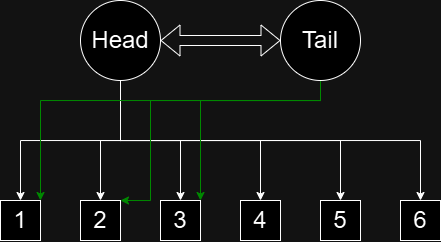
**Bioinformatic assignment: Hidden Markov models**

1. **Hidden Markov Games**

Imagine the following game. You have a coin and a dice. In each row you flip the coin and throw the dice. If the coin shows head, you record the number of the dice, if the coin shows tail, you record the rounded number of the dice (1,2 will be recorded as 1; 3,4 will be recorded as 2 and 5,6 as 3).

* 1. Draw the model of this Hidden Markov model, what is the hidden what is the observed state?
  2. Considering the following chain of results:

1 2 3 2 4 5 3 3 1 6 ; for each step, what is the probability that the coin was showing heads or tails, respectively?

1. **Hidden Markov for transmembrane domains**One popular application of hidden Markov models is the prediction of transmembrane domains. Predict the trans-membranes domains in the HXT6 protein.
   1. How many transmembrane domains were predicted and where are they?
   2. Looking at the Transmembrane domains, what is the length distribution of the domains, and what is the fraction of hydrophobic residues inside those domains compared to outside.
   3. Considering the second TM helix (117-134), which amino acid would you need to change to perturb the prediction. Give one example for the substituted sequence,
   4. It also seems like the length of the potential TM is important for the prediction. Try removing residues from the 2nd TM until the prediction is lost, what is the minimum number of residues you need for a helix?
2. **Hidden Markov for protein domain prediction.**Previously you analyzed different human caspase paralogs. Now you can use protein domain prediction to better understand the similarities and differences between the different caspase paralogs. Determine the domain structure for all 5 caspase paralogs (1,2,3,4 and 5).
   1. Which domains do you find?
   2. How does this compare to the sequence alignment / the phylogenetic tree of the caspases?
3. **Generating your own HMM model**The PFAM HMM models are originally created from MSA of known members o a particular domain. Similarly, you can create your own domain and scan proteomes for the occurrence of this domain.
   1. Given that you want to generate a general CASPASE domain, which parts of the alignment would you use?
   2. Generate a HMM model from the alignment and use it to find “caspases” in Zebrafish.
   3. How many “caspases” did you find?
   4. Compare the HMM profile to the MSA what happened to highly conserved residues? What happened to less conserved residues?
   5. Both the Catalytic Cystein (Cys 163 in Casp-3) and the adjacent alanine (Ala-162) are perfectly conserved, why do you think their height is different?
   6. Lastly, try using only three Caspases (1,4,5) for the Domain generation how many what are the effects on 1) the number of “detected caspases”, 2) the e-values and the 3) the HMM profile.
4. 
   1. The hidden states are the result of the coin toss, the observed states are the recorded dice results
   2. 1/3 head – 2/3 tail for 1,2,3 as recorded result, for 4,5,6 it’s head with a 100% probability
5. –
   1. There are 12 transmembrane domains:
      1. 10-30
      2. 50-70
      3. 90-110
      4. 130-150
      5. 170-190
      6. 210-230
      7. 250-270
      8. 290-310
      9. 330-350
      10. 370-390
      11. 410-430
      12. 450-470
   2. The length is usually 20 ammino acids, to calculate the fraction I use tm2:
      1. Sequence: SIMCIMIAFGGFVFGWDTGT
      2. Hydrophobic residues: S,I,M,C,I,M,I,A,F,G,F,V,F,G,W (15 out of 20)
      3. 75% of all ammino acids are hydrophobic residues
   3. By changing Leucine (L) at position 120 to Glutamic acid (E) I substitute a hydrophobic acid to a hydrophilic one:
      1. Original: SIMCIMIAFGGFVFGWDTGT Substituted: SIMCIMIAFGGFVFGWDTGT
   4. The minimum number of residues required for the helix is around 15, reducing further loses the prediction.
6. –
   1. The domains found are the following:
      1. CARD domain
      2. Peptidase\_C14 domain

Except in casp 3 where only the Peptidase\_C14 domain is found.

* + 1. On the sequence alignment the Peptidase\_C14 domain is highly conserved among the caspases too, as in the protein prediction.
    2. Using a phylogenetic tree the following caspases: caspase-1, -4, and -5, are more closely related

1. –
   1. The peptidase\_C14 domain is highly conserved so it should be included in the HMM model.
   2. –
   3. The most present caspases were casp 1, 4 and 13
   4. In the HMM profile, highly conserved residues will have high emission probabilities and will be represented with high scores in the profile, while less conserved residues will have lower emission probabilities and lower scores in the HMM profile.
   5. The different height can be due to the different background frequencies of these amino acids. Cysteine is less common in proteins compared to alanine, so its conservation is more significant, leading to a higher score.
   6. Using fewer sequences will reduce the sensitivity of the HMM, potentially leading to fewer detected caspases, the e-values might increase and the hmm profile might be less robust