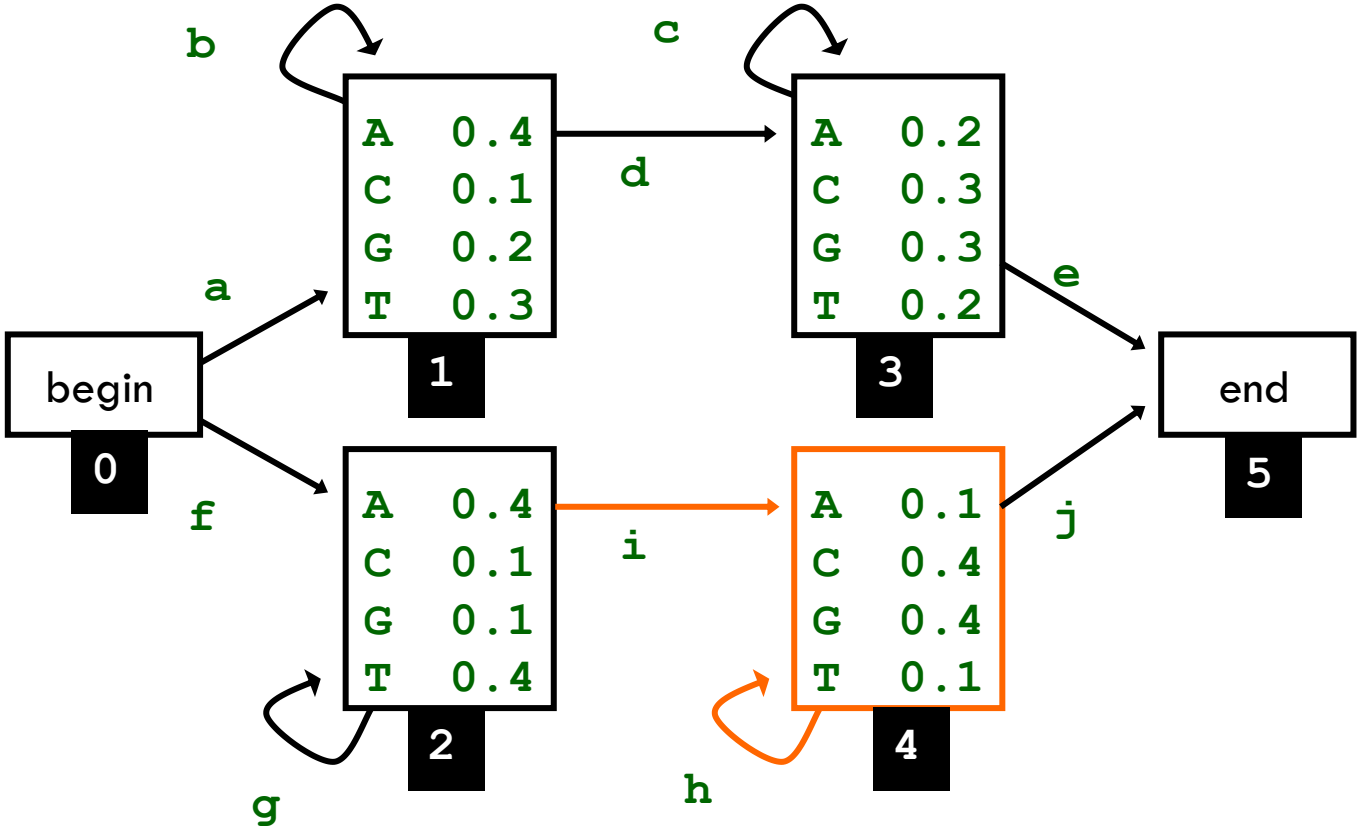
BIOINFORMATIKA – CV. 5
HRÁTKY S HMM

Skryté Markovské modely – jednoduché experimenty

Než začneme... učení se parametrů pomocí Baum-Welchova algoritmu

- Máme k dispozici strukturu modelu a (dlouhou) sekvenci znaků (nikoliv však stavů) vygenerovanou tímto modelem. **Jak určíme jeho parametry?**



Baum-Welchův algoritmus

- Máme k dispozici strukturu modelu a (dlouhou) sekvenci znaků (nikoliv však stavů) vygenerovanou tímto modelem. **Jak určíme jeho parametry?**
- Nemůžeme použít jednoduchý přístup pro odhad přechodových pravděpodobností a pravděpodobností emise, protože **nemáme k dispozici posloupnost stavů!**
- **Řešení: varianta EM algoritmu** (Baum-Welchův algoritmus)
  - Iterativní algoritmus
  - Nalezení globálního optima není zaručeno
  - Často je důležitý počáteční odhad parametrů

Experiment 1: rozpoznávání prokaryotických a eukaryotických DNA sekvencí pomocí **HMM**

- K dispozici máte 4 stejně dlouhé DNA sekvence. Dvě pocházejí z prokaryotních a dvě z eukaryotních organismů (délky genomů se typicky liší, ale tím bychom si problém příliš zjednodušovali).
- Navrhněte strukturu HMM a počáteční odhad parametrů pro tento model. Použijte tento počáteční odhad pro odhad parametrů pomocí Baum-Welchova algoritmu (matlab: **hmmtrain**) zvlášť pro sekvence prokaryotních a zvlášť pro sekvence eukaryotních organismů.
- Na základě takto natrénovaných HMM rozhodněte, který ze dvou souborů označených jako *unknown1.mat* a *unknown2.m* obsahuje sekvenci pocházející z prokaryotního a který z eukaryotního organismu. (Pro výpočet logaritmu pravděpodobnosti sekvence použijte funkci **hmmdecode**).

- **K čemu jste dospěli? Zkuste najít nějaké biologické zdůvodnění** (například na Wikipedii)

Experiment 2: „gene finding“ pomocí **HMM**

- Semestrálka?