Markovovy modely v Bioinformatice

Outline

- Markovovy modely obecně
- Profilové HMM
- Další použití HMM v Bioinformatice

Analýza biologických sekvencí

- Biologické sekvence: DNA,RNA,protein prim.str.
- Sekvenování snadné vs. např. 3D-struktura
- Velké množství experimentálních dat
- Cíl: „Využít tyto data k zjištění užitečných informací o neznámých sekvencích.“

Analýza biologických sekvencí

- Dva základní přístupy:
  - Data mining
    - Přímo v datech se hledají zajímavé informace
  - Strojové učení
    - Pomocí učících dat se vytvářejí modely které analyzují neznámá data

Metody strojového učení

- (skryté) Markovovy modely
- Neural networks
- Support Vector Machine
- další: Genetic algorithms,Bays networks…
- Každá metoda má své výhody

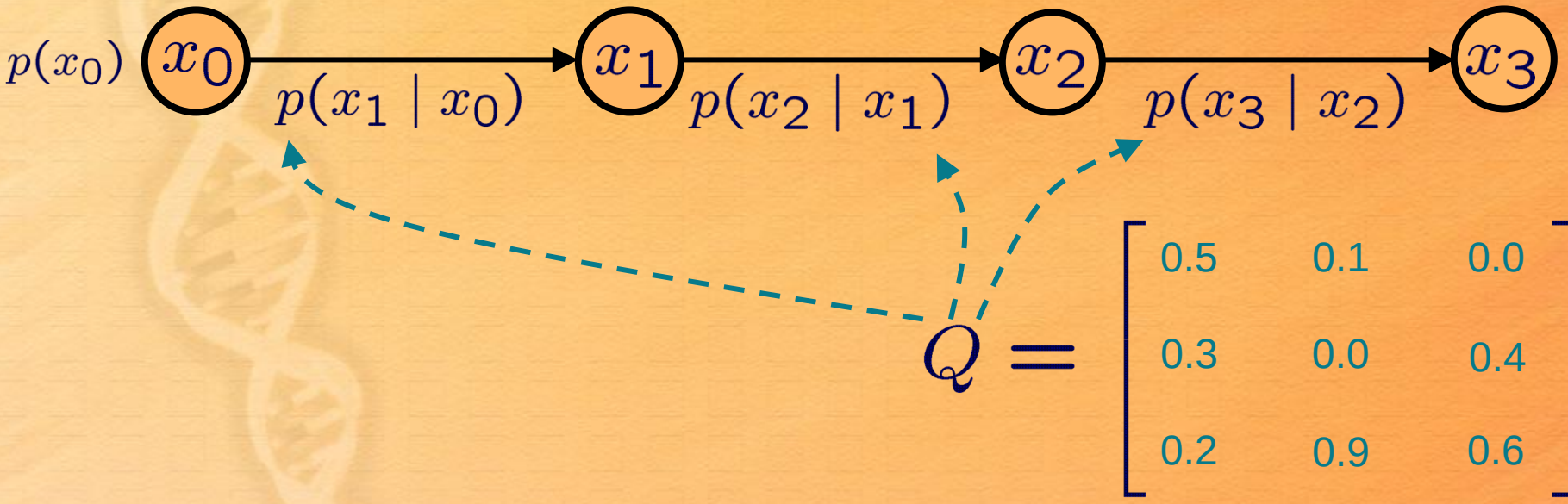
Stochastický (náhodný) proces

- Množina náhodných proměnných indexovaných v čase
- Nejjednodušší případ: diskrétní proces
  - Sekvence náhodných veličin

  ! Pro naše potřeby stačí diskrétní veličiny
  - Sekvence znaků generovaných náhodně ale s definovanou pravděpodobností

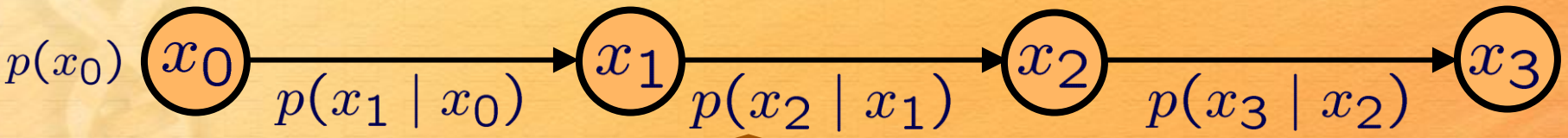
Markovův řetězec

- Diskrétní stochastický proces s vlastností „bezpamětnosti“ (Markov property)
  - Pravděpodobnost přechodu ze stavu e_i do stavu e_j nezávisí na tom, jak se systém do stavu e_j dostal.



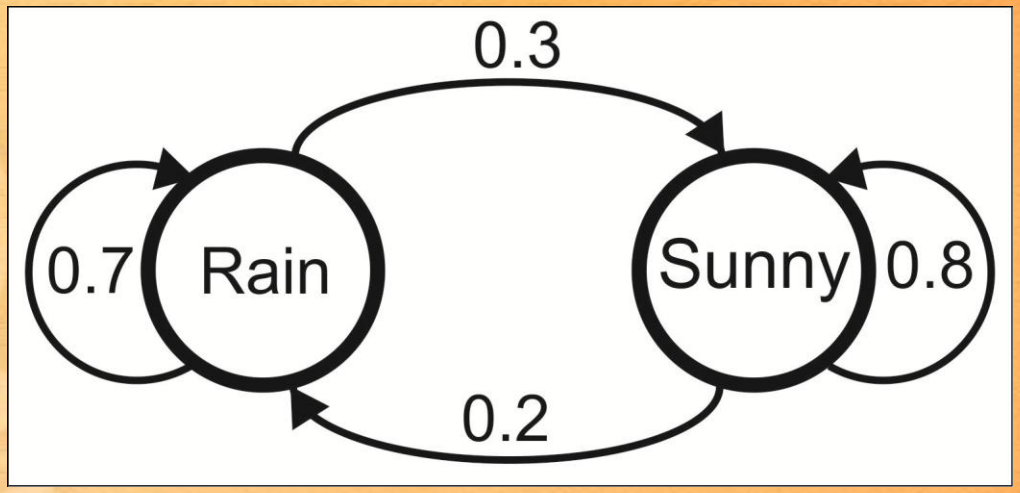
Markovův řetězec vs. Markovův model

- Totožné pojmy



model generuje řetězec

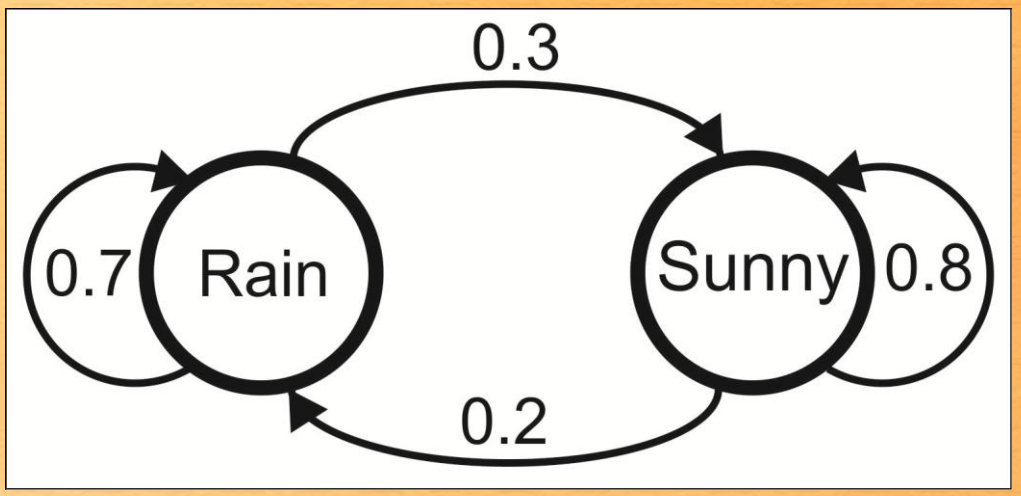
řetězec definuje model



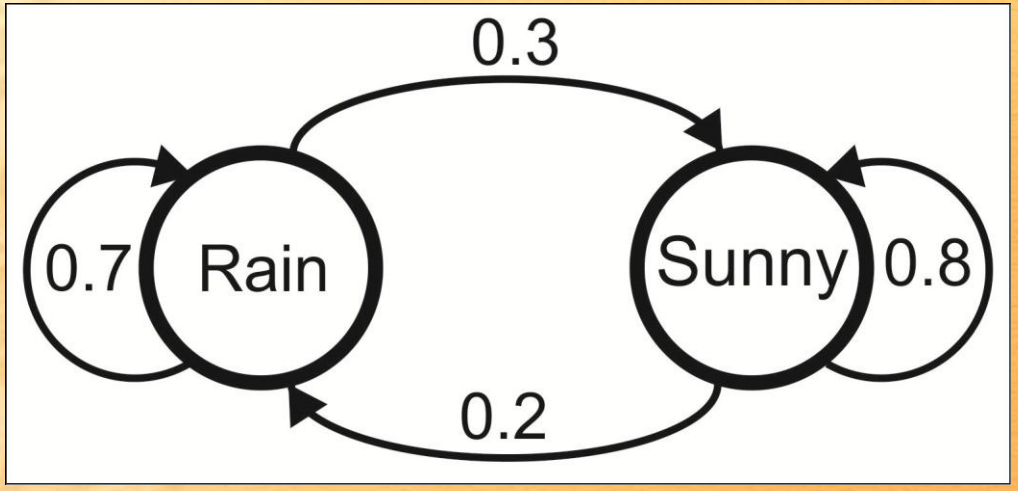
Markovův model vs. Sekvence znaků

RRRSSSSRRSSRRSS

Můžeme určit pravděpodobnost sekvence vůči modelu



Markovův model

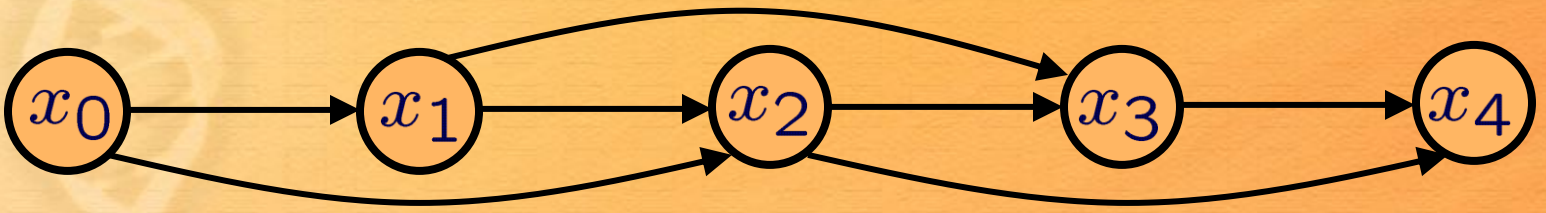


- Definuje se pomocí tranziční matice:

$$Q = \begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{bmatrix}$$

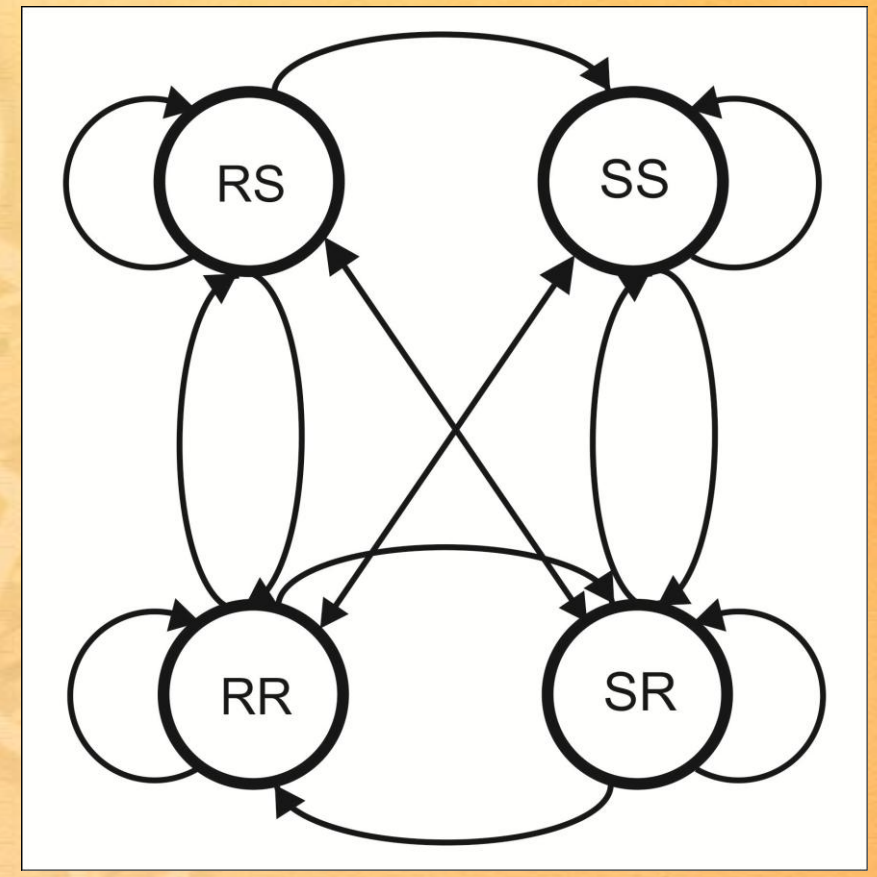
Markovovy modely vyššího řádu

- Markovův model n-tého řádu
  - Pravděpodobnost přechodu ze stavu e_i do stavu e_j závisí na n předchozích stavech



Markovovy modely vyššího řádu

- Lze převést na standardní HMM



Markovovy modely vyššího řádu

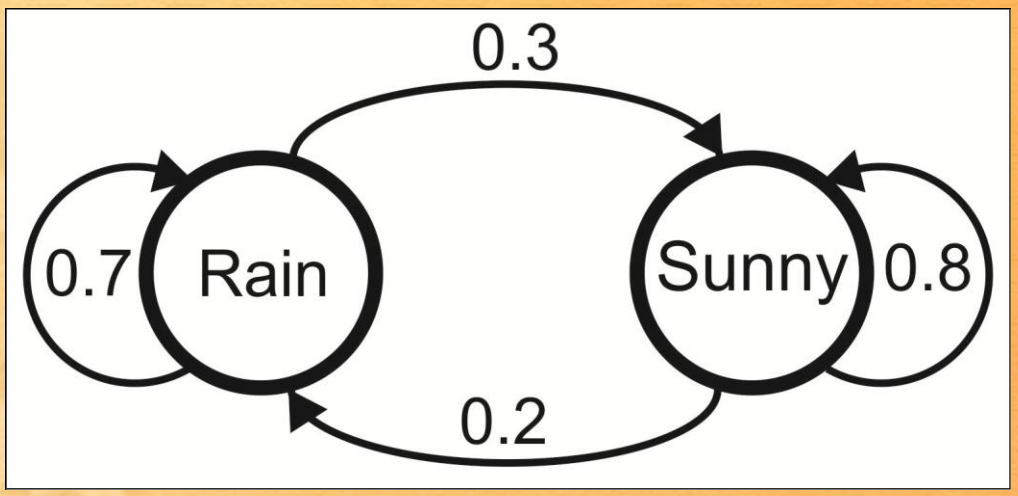
- Ukázka z lingvistiky
  - Order-1: Ched t ainone wand LORD, Thenathan g u t; w t Sona t essose Anasesed an trer.
  - Order-2: a grand it the woraelfspit up fir as king thaderld, I slot kins ts,
  - Order-3: Against of Ashekeloverth with his uncill be pill prehold, We came womb?
  - Order-4: Open he sister daughteousness, whitherefore they present with themselves;
  - Order-5: Thus saith thy man righted, behold, Gaal was thou art that fell do them
- http://www.yisongyue.com/shaney/

Využití v bioinformatice

- Nepříliš využitelné při analýze sekvencí
  - Předpovídání promotorů
- Semi-Markov modely v systémové biologii
- Základ pro skryté Markovovy modely

Skryté Markovovy modely (HMM)

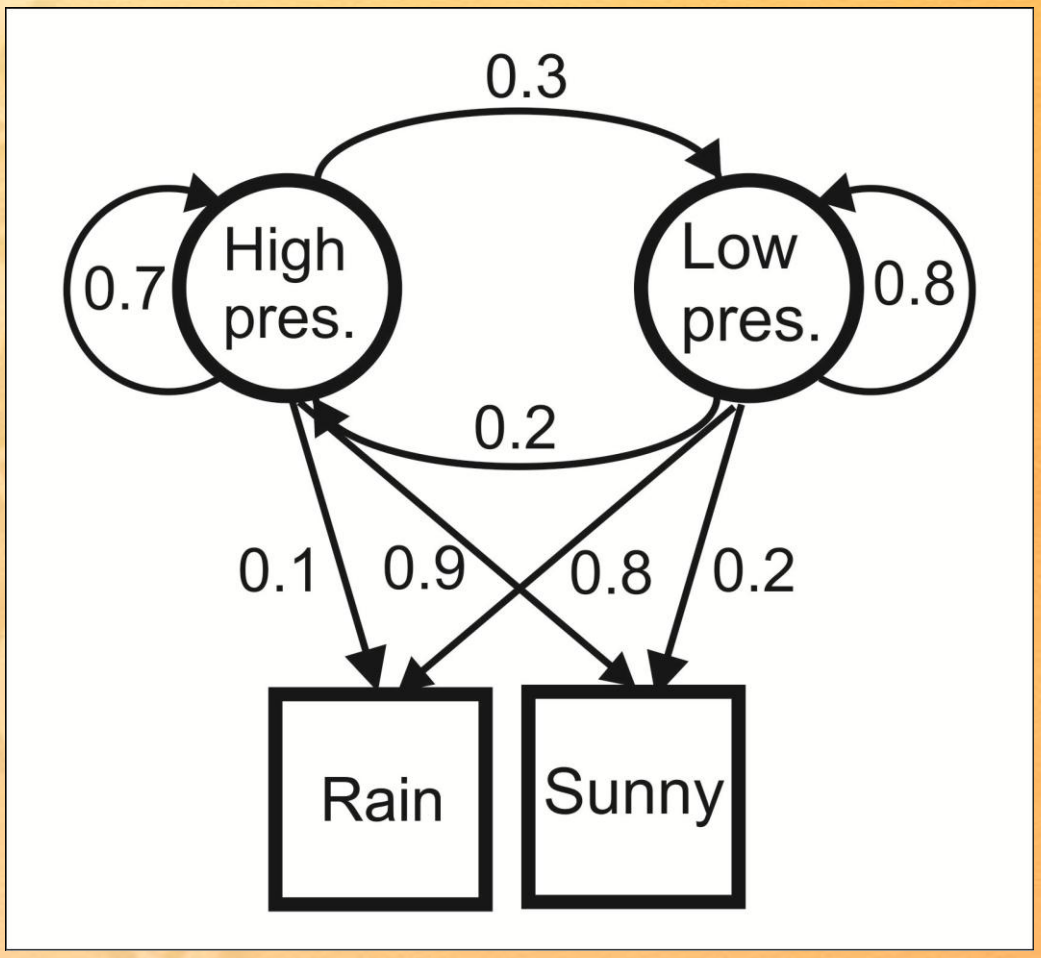
- Doposud každý stav reprezentoval konkrétní znak v řetězci



RRRSSSSRRSSRRSS

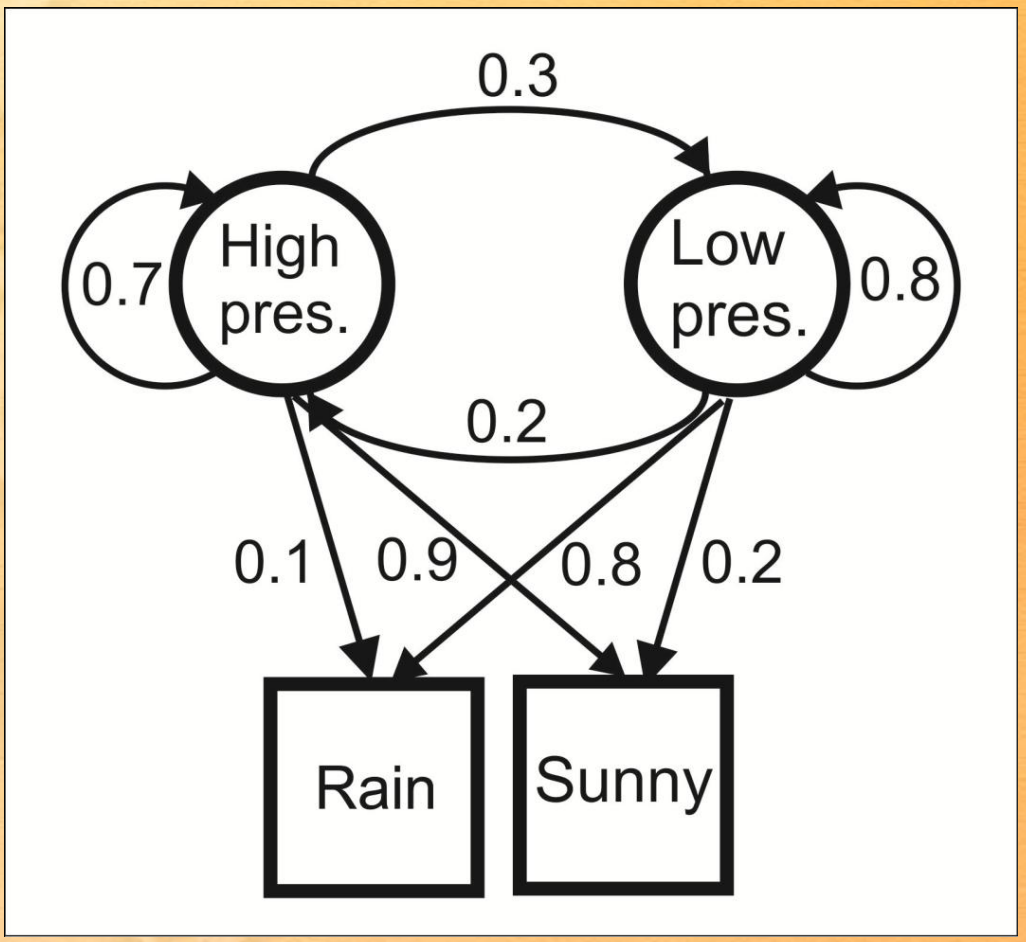
HMM

- Mějme stavy které mohou generovat různé znaky řetězce s definovanou pravděpodobností



HMM

- Tyto stavy jsou skryté vydíme pouze pozorování které generují

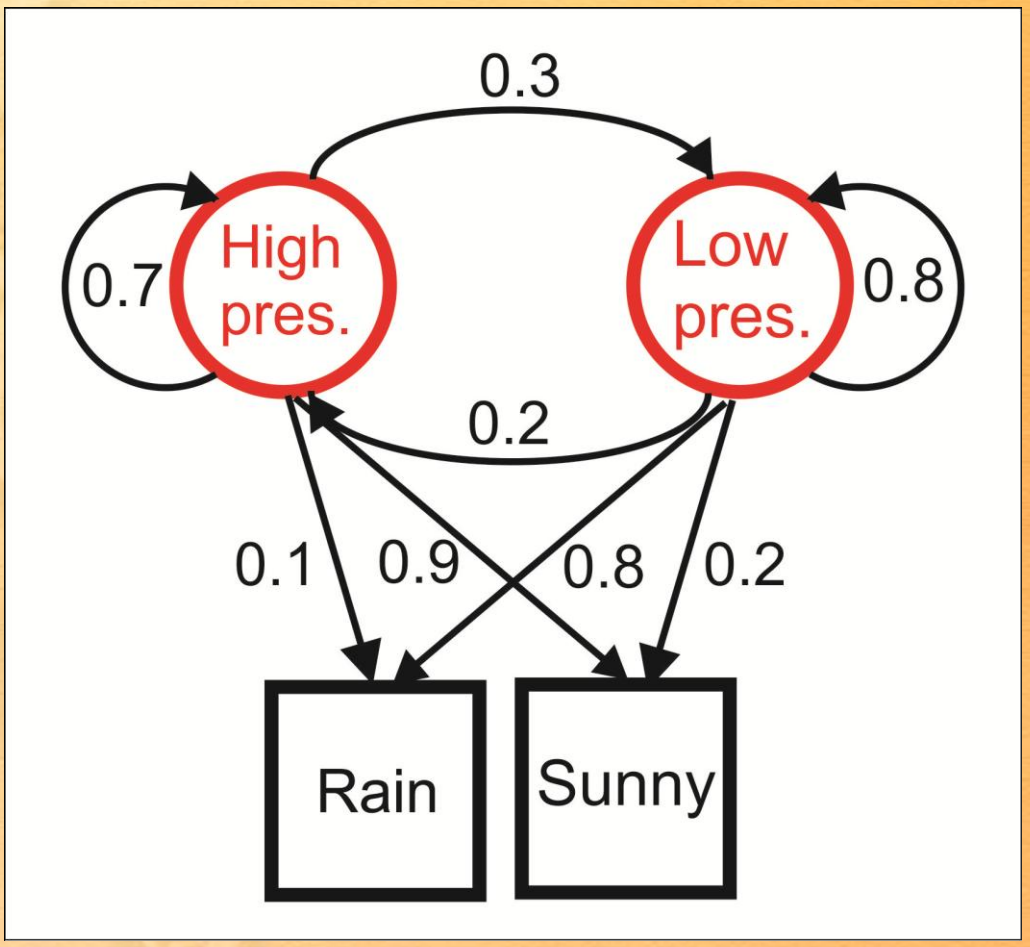


Definice HMM

- K definování potřeba 2 matice a 1 vektor
  - Tranziční matice přechodů mezi skrytými stavy
  - Matice pravděpodobností generovaní jednotlivých pozorování danými stavy
  - Vektor rozložení pravděpodobnosti počátečního stavu

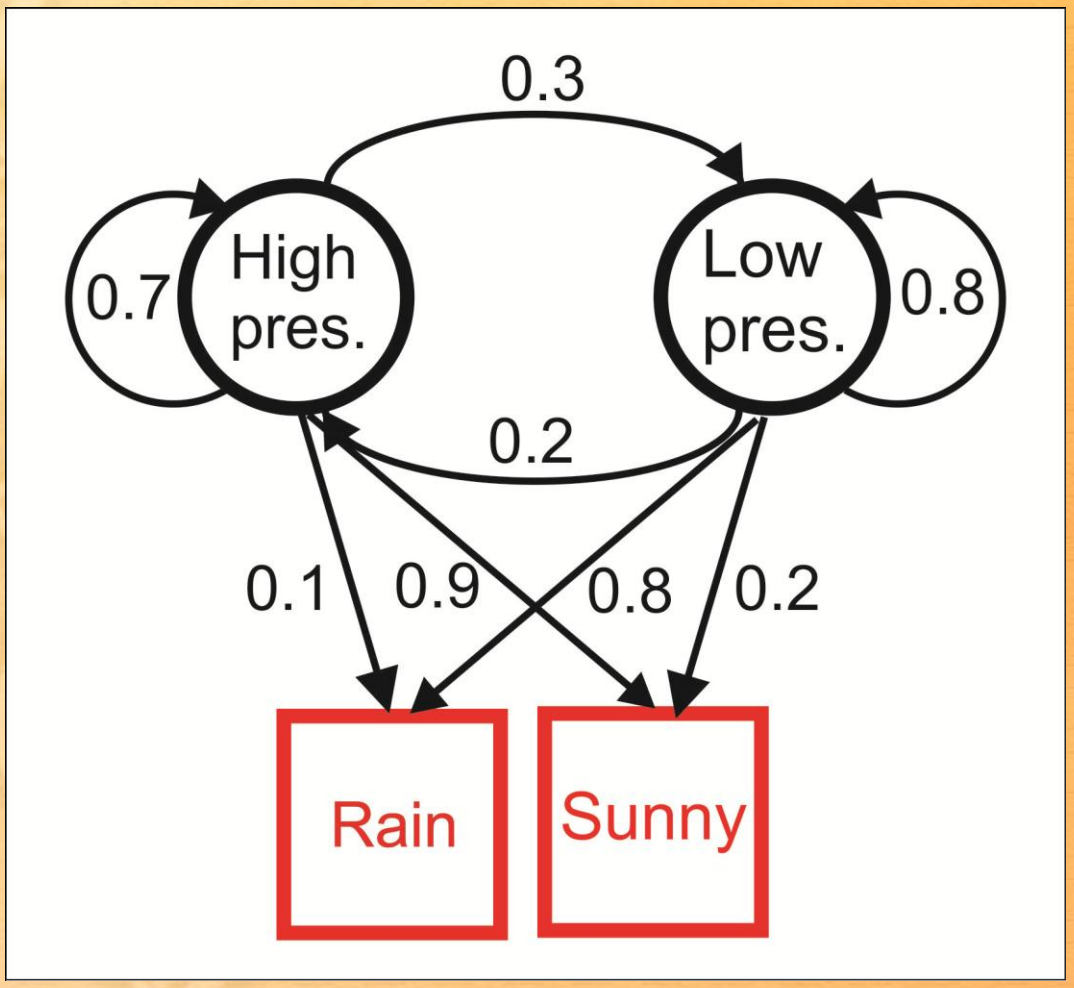
HMM

- Skryté stavy



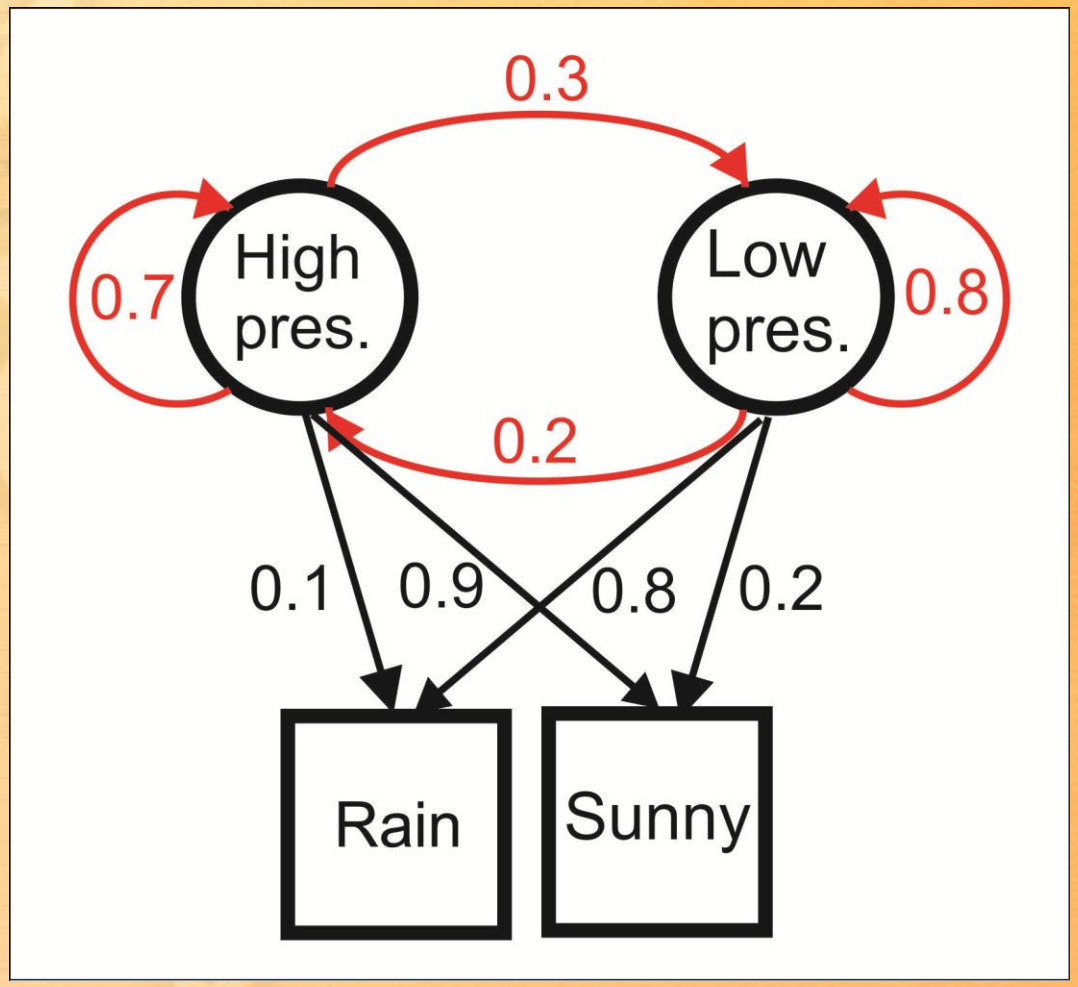
HMM

- Možná pozorování



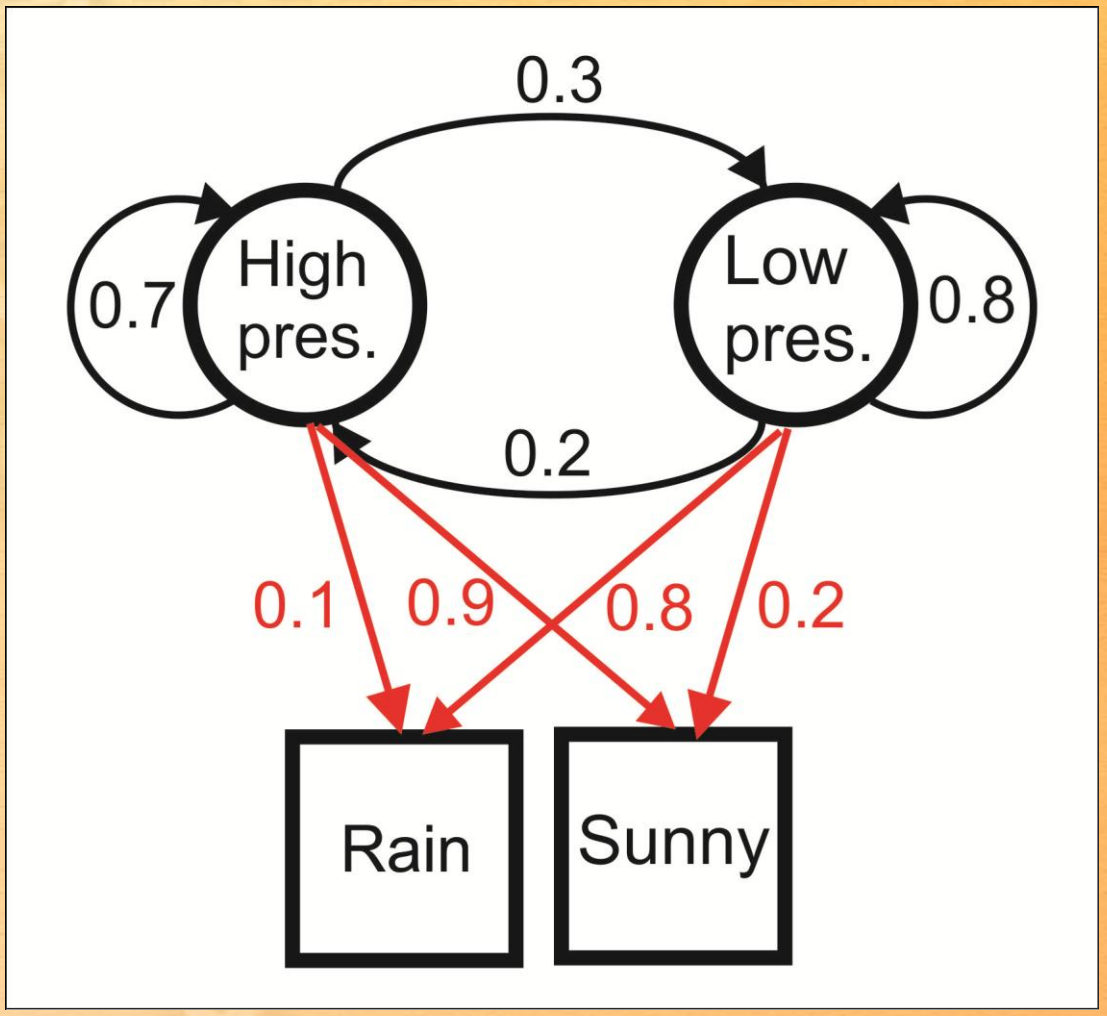
HMM

- Tranziční pravděpodobnosti



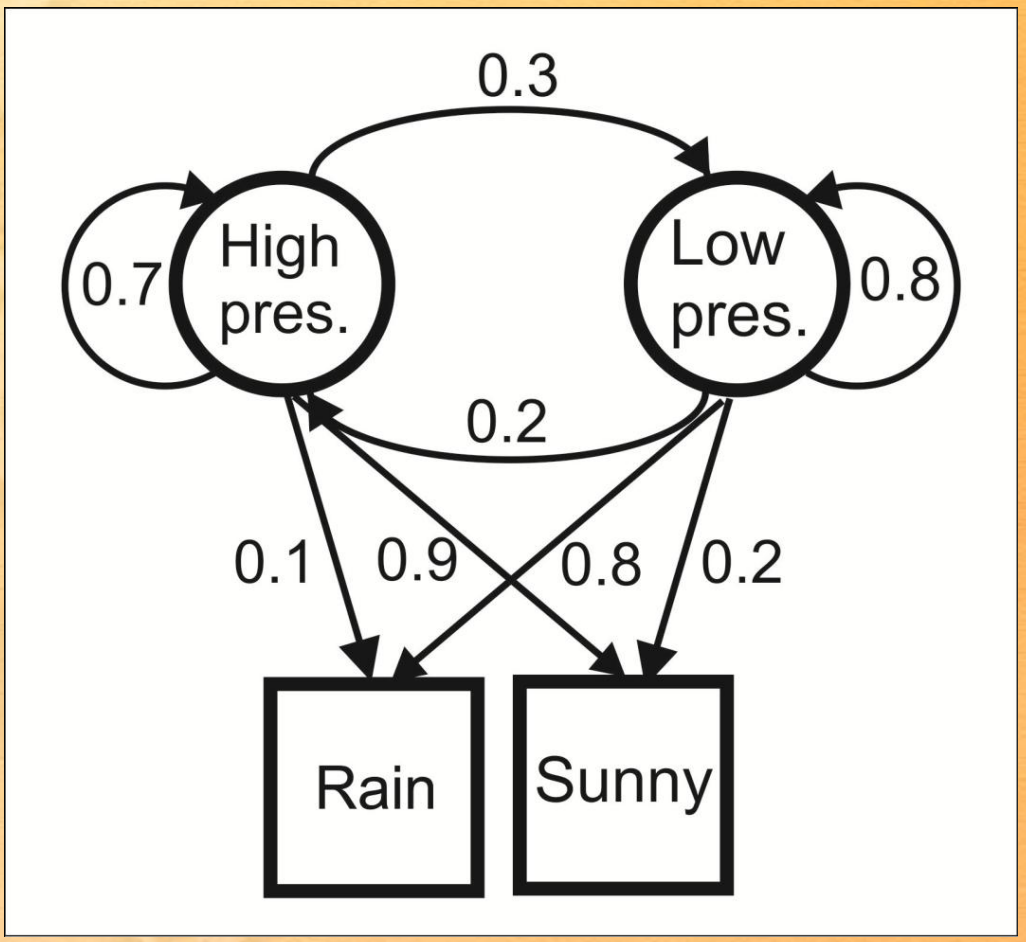
HMM

- Emisní pravděpodobnosti

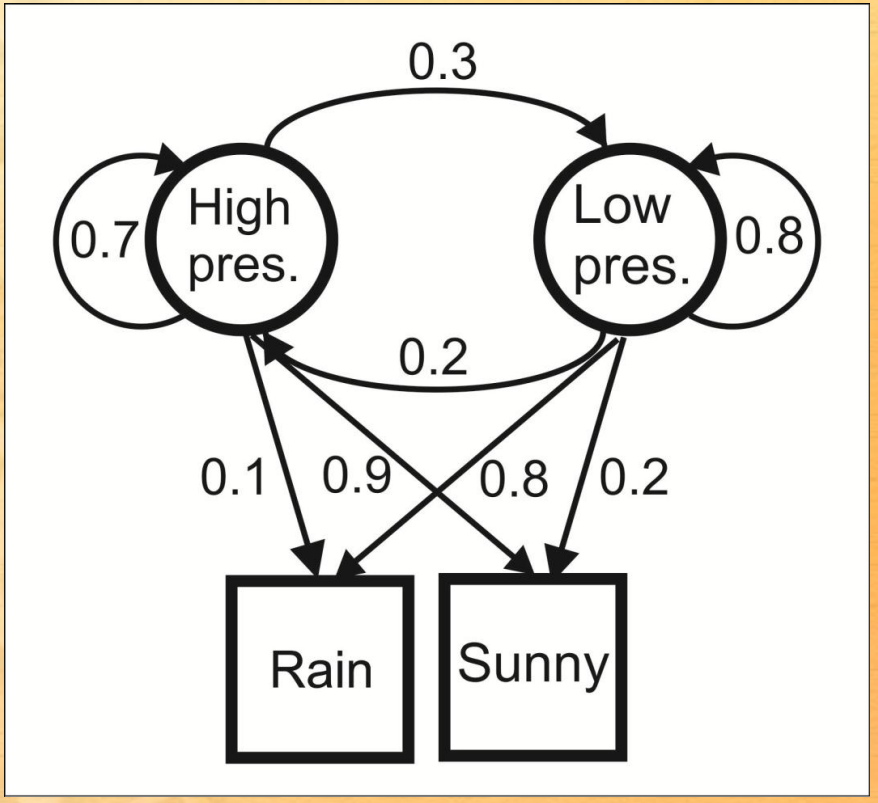


HMM

- Počáteční stav = (0,1)



HMM



- Mějme pozorování

S1: RSRS
S2: RRSS

Co můžeme chtít zjistit?

Co můžeme chtít zjistit?

- S jakou pravděpodobností generuje model danou sekvenci

Co můžeme chtít zjistit?

- S jakou pravděpodobností generuje model danou sekvenci?
- Jaká je nejpravděpodobnější průchod skrytými stavy při generování dané sekvence?

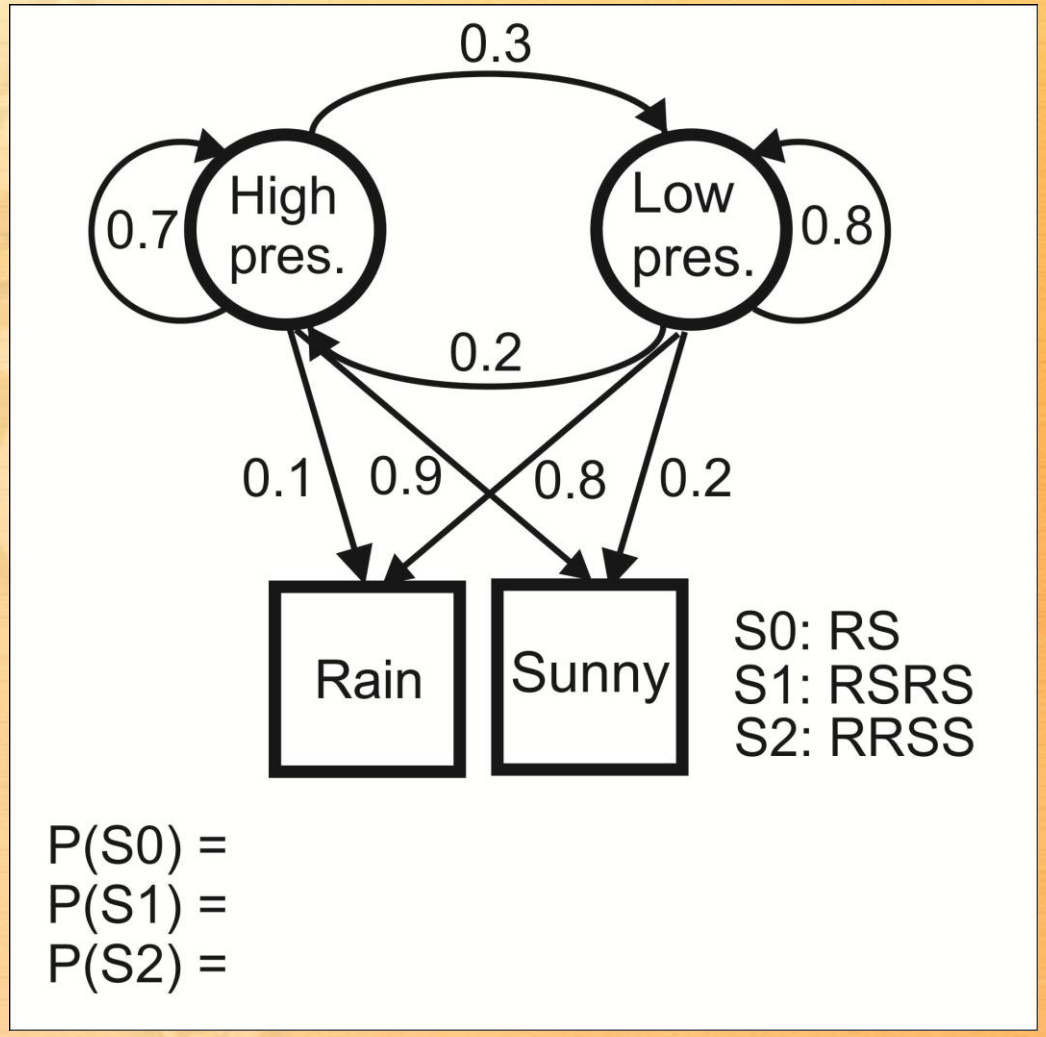
Co můžeme chtít zjistit?

- S jakou pravděpodobností generuje model danou sekvenci (evaluace)
- Jaká je nejpravděpodobnější průchod skrytými stavy při generování dané sekvence (dekódování)
- Jak změnit parametry modelu tak aby generoval danou sekvenci s větší pravděpodobností? (učení)

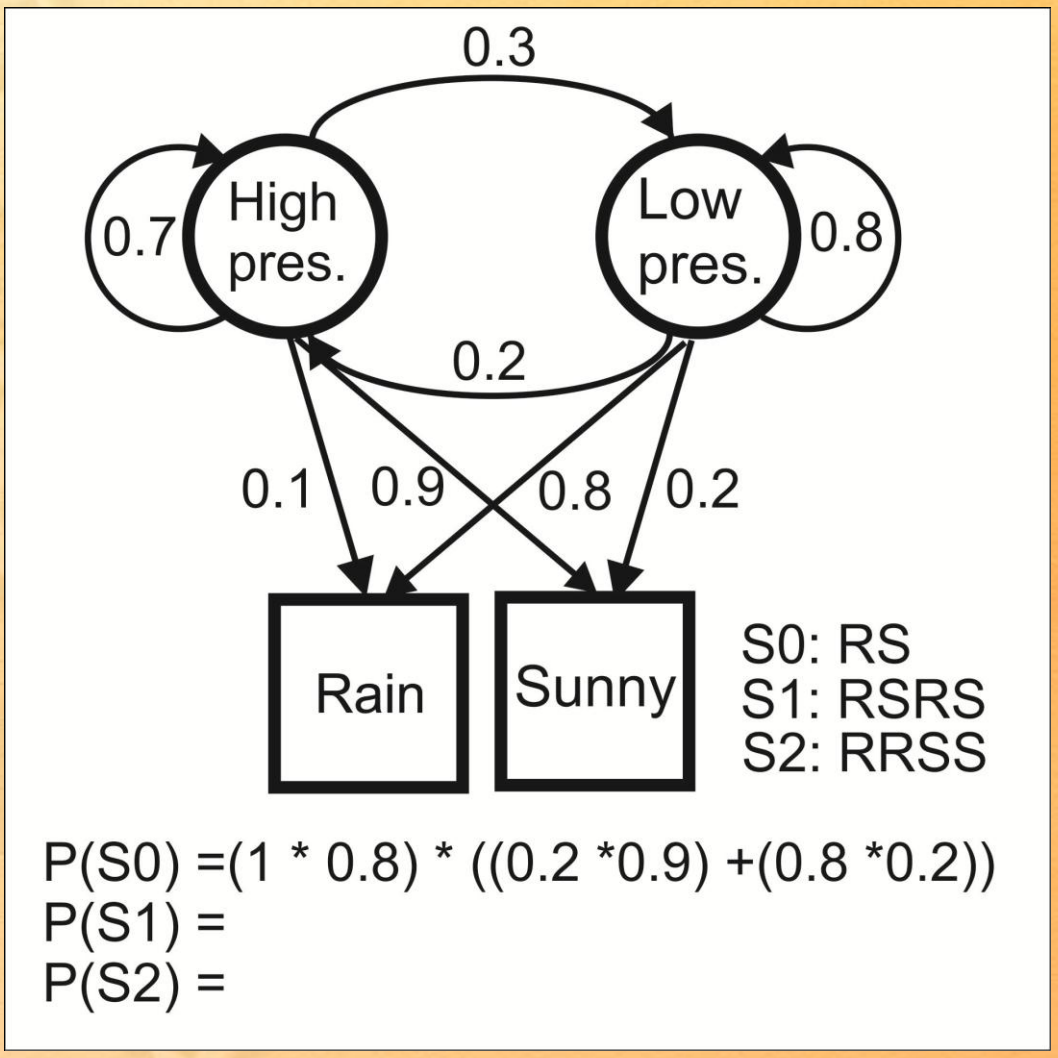
Algoritmy

- Evaluace: Forward-Backward algorithm
- Dekódování: Viterbi algorithm
  - Algoritmy dynamického programování
  - Oba optimální
  - Složitost $O(P*S^2)$
- Učení: Baum-Welsch algorithm
  - Zasekává v lokálních maximech => potřeba solidní odhad prvního modelu
  - Model je jen tak dobrý jak učící data

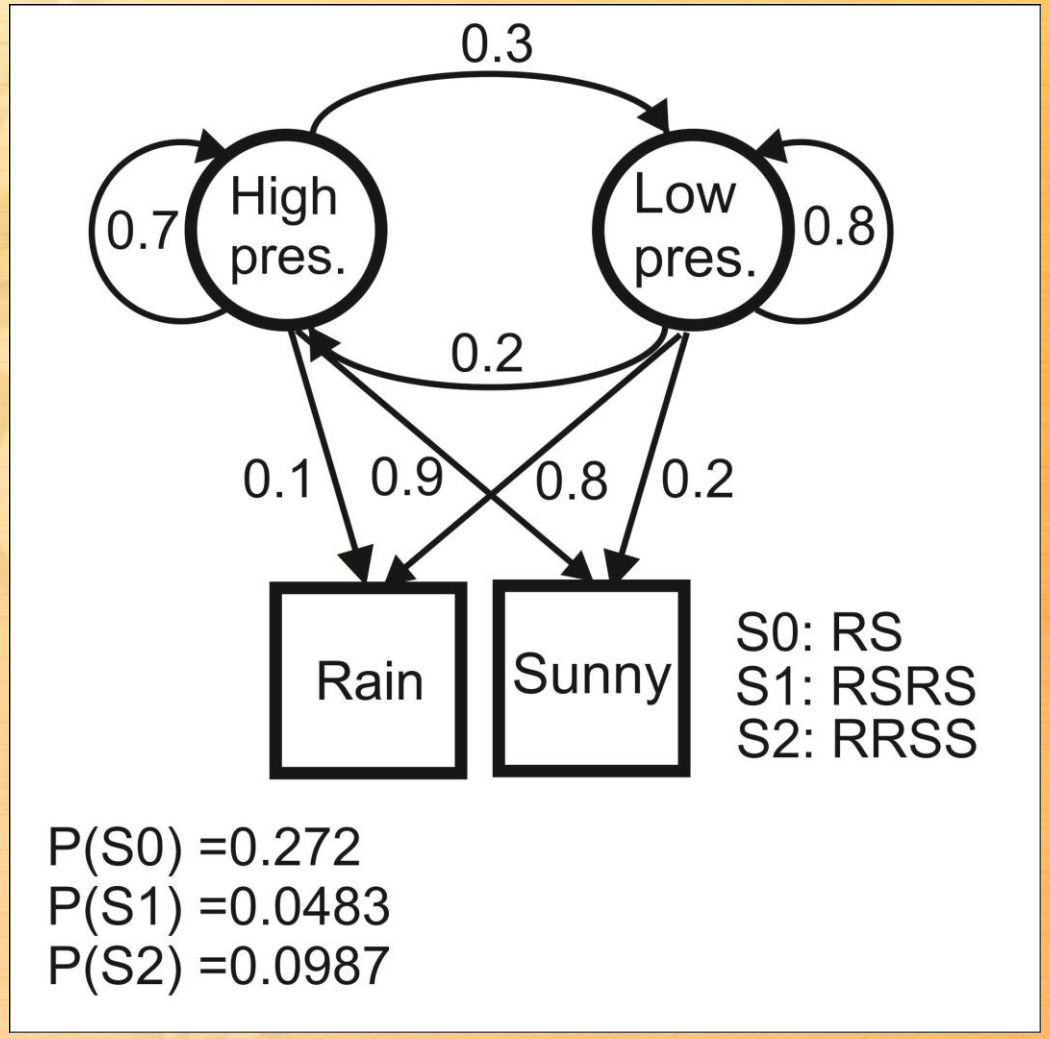
Evaluace: Forward-Backward algorithm



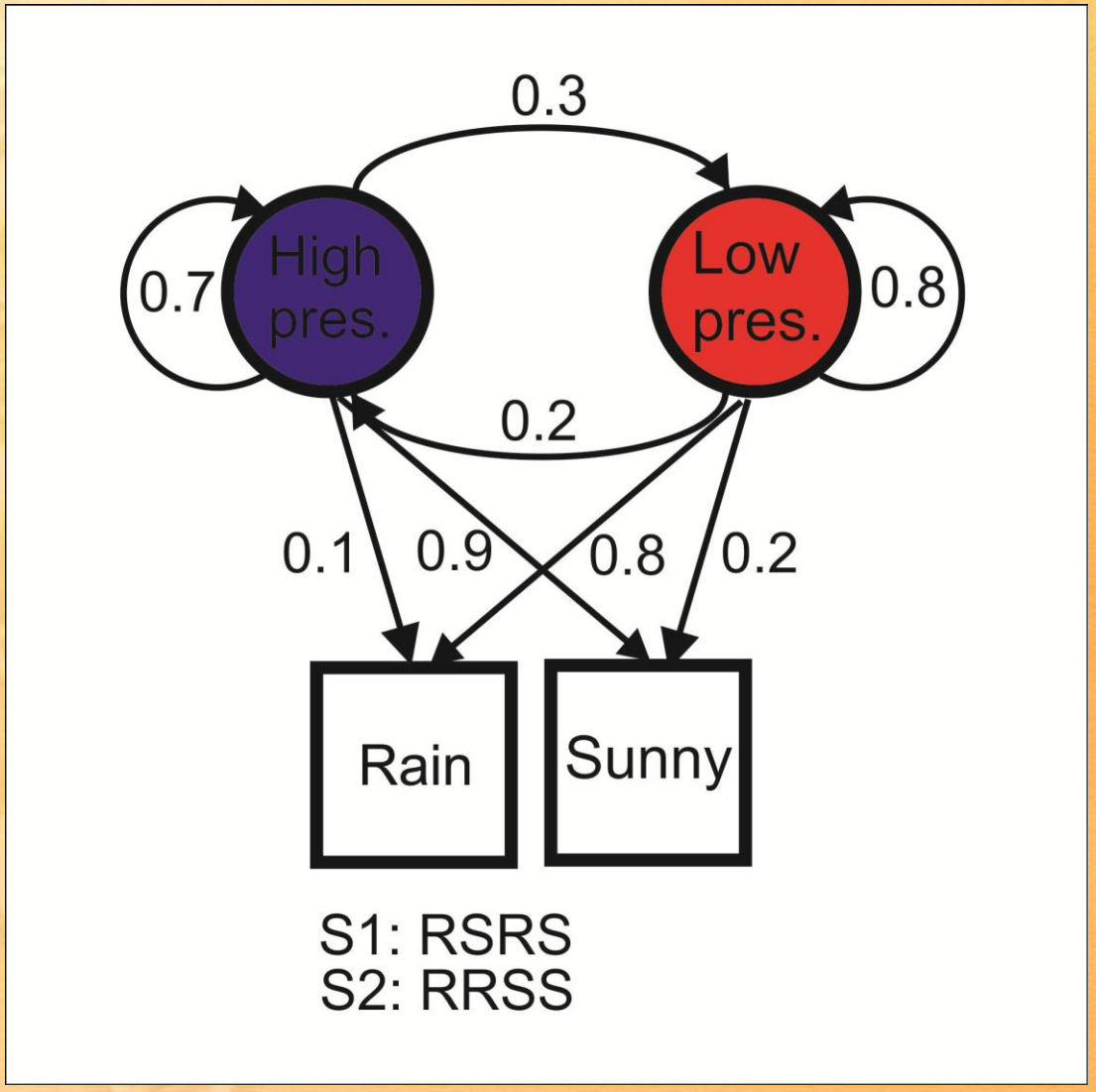
Evaluace: Forward-Backward algorithm



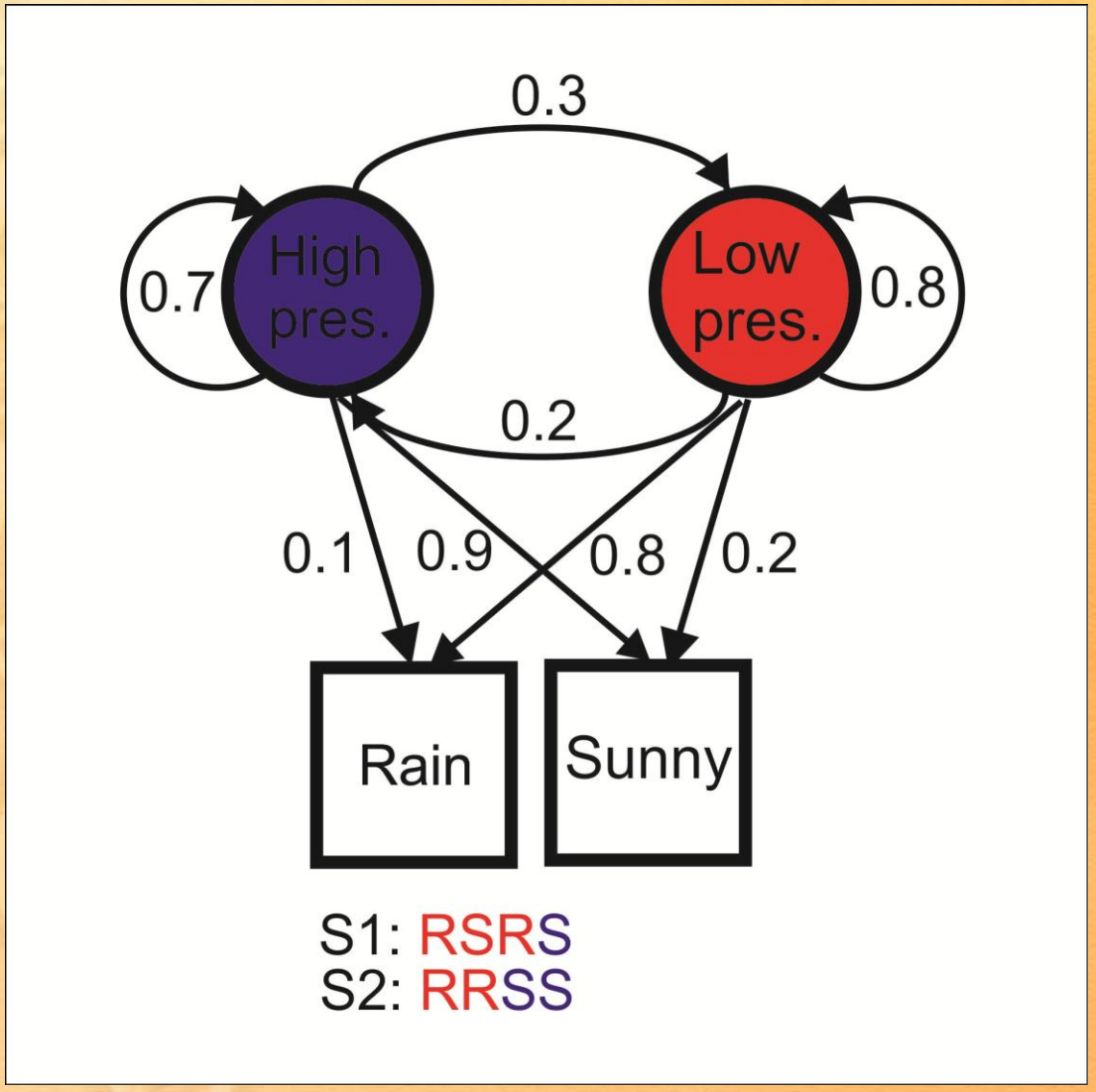
Evaluace: Forward-Backward algorithm



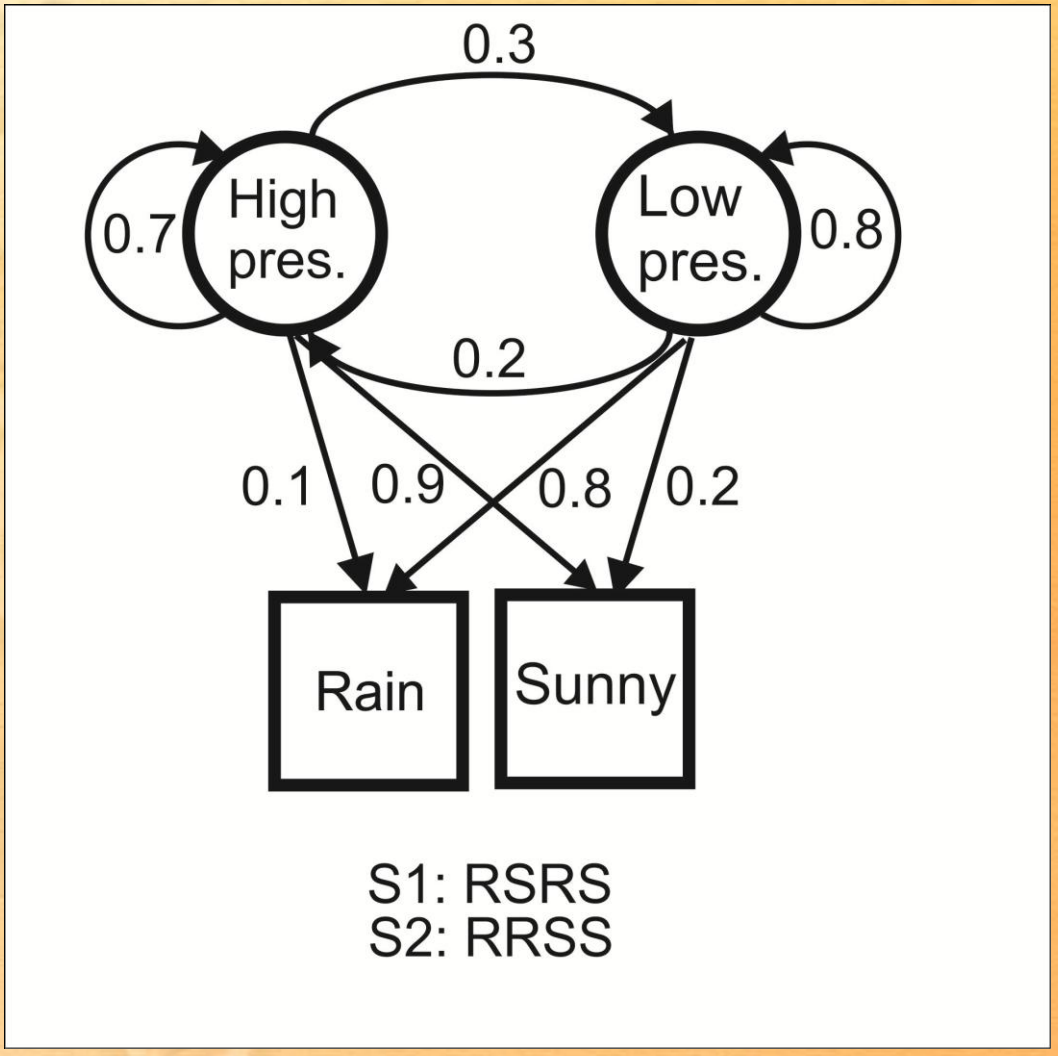
Dekódování: Viterbi algorithm



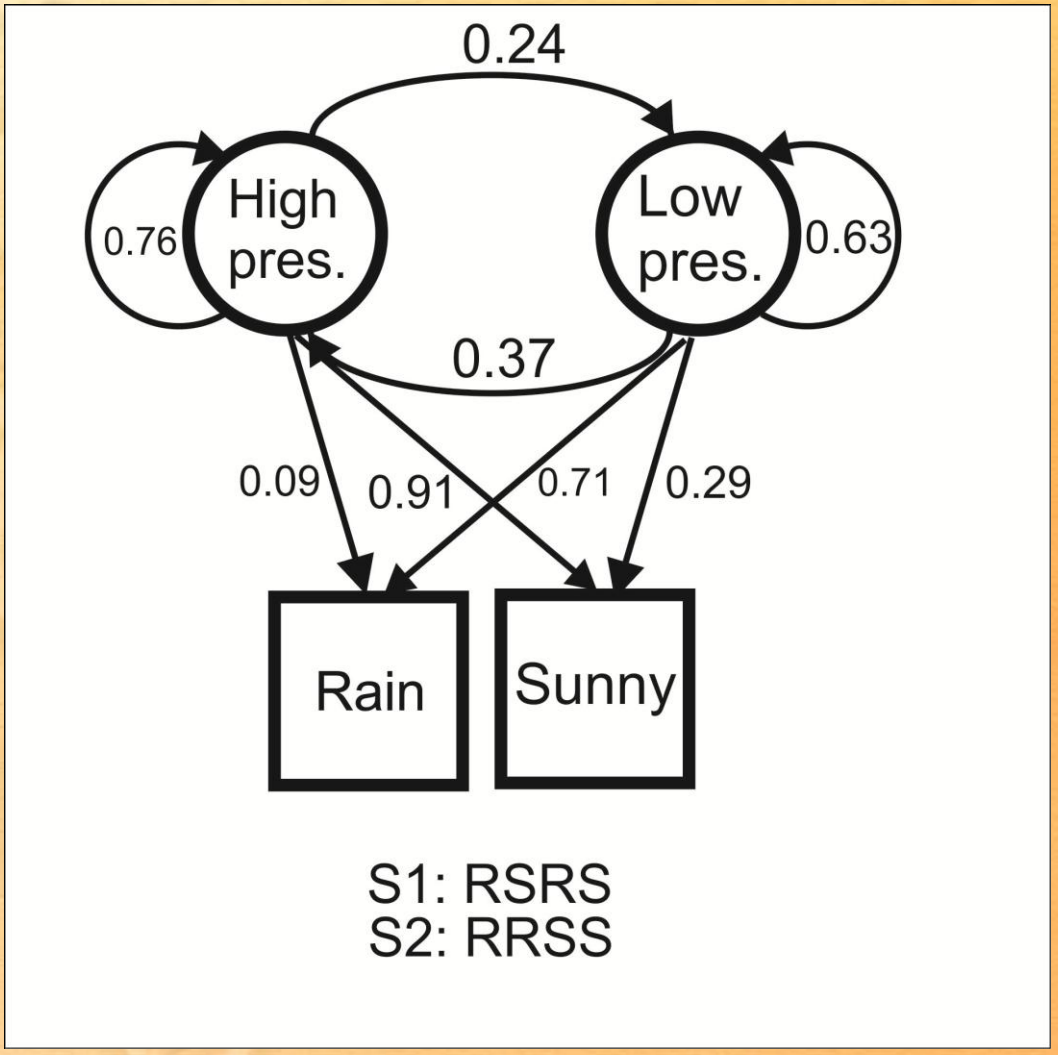
Dekódování: Viterbi algorithm



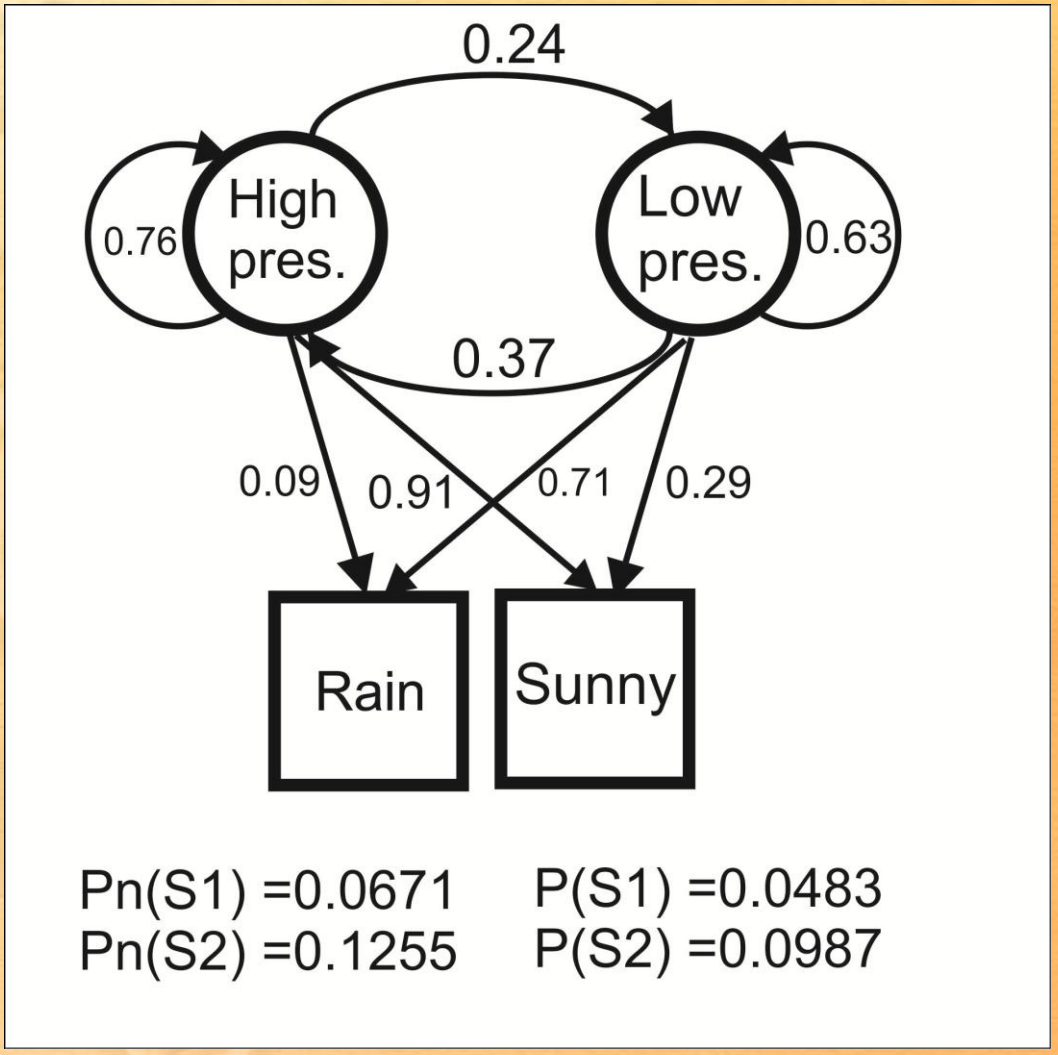
Učení: Baum-Welsch algorithm



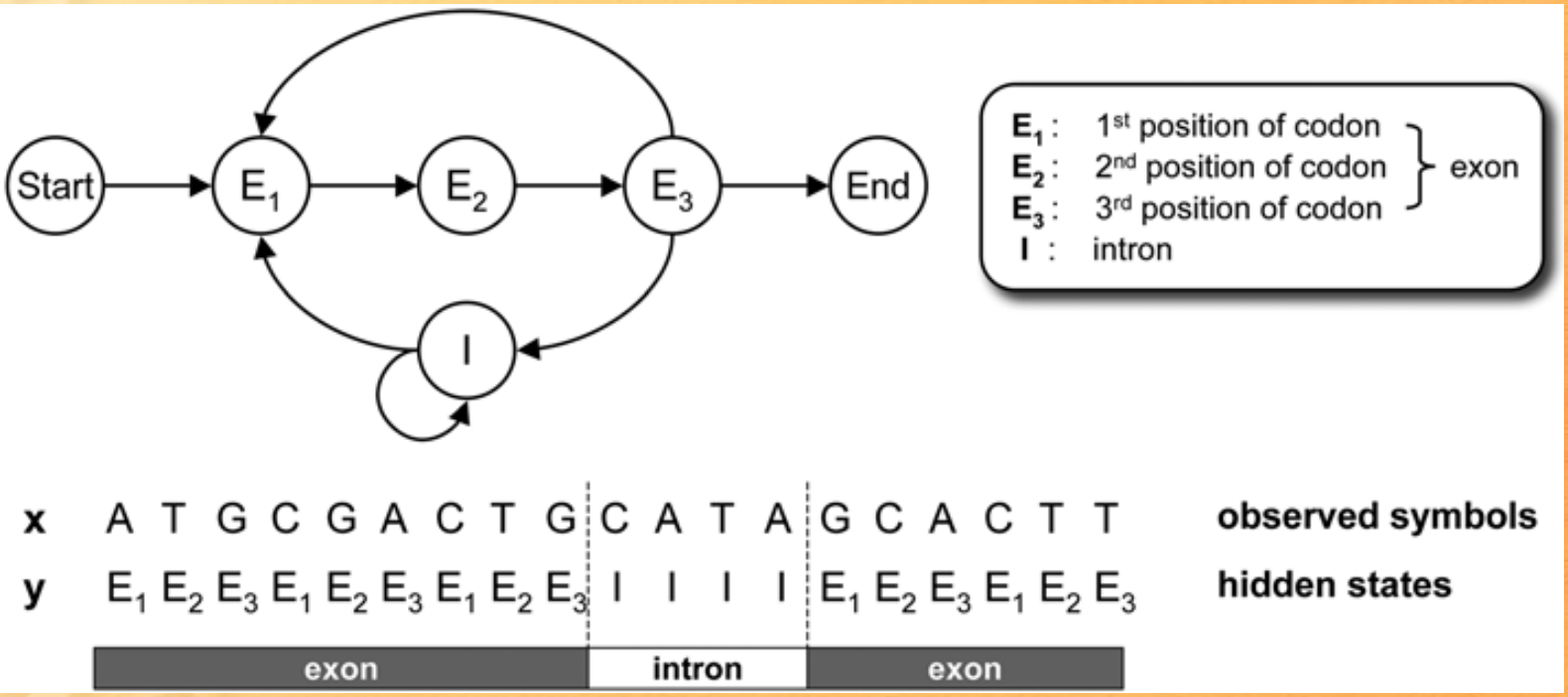
Učení: Baum-Welsch algorithm



Učení: Baum-Welsch algorithm



Jednoduchý příklad z bioinformatiky



Pouziti

- Obecně jakýkoliv klasifikační problém
  - v rámci sekvence
    - Hledání genů
    - Předpověď sekundární struktury proteinů
    - Předpověď domén proteinů
    - Struktura RNA
  - Celé sekvence
    - Klasifikace do proteinů do rodin
    - Profilové HMM

Příště

- Markovovy modely obecně
- Profilové HMM
- Další použití HMM v Bioinformatice