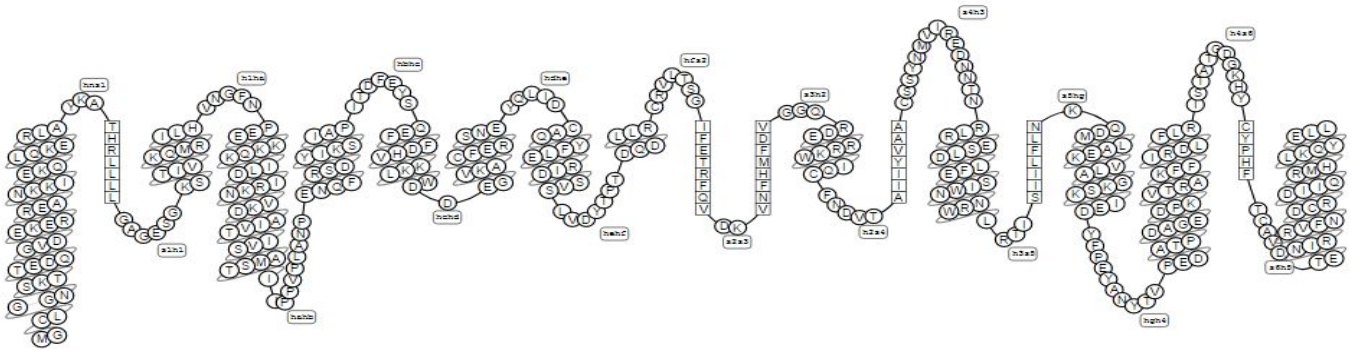
# Predikce topologie transmembránových proteinů

Šimon Rozsíval

# Topologie transmembránových proteinů

- Přesnou zjištěnou strukturu proteinů experimentálně máme k dispozici jen pro část proteinů
  - Experimentální zjištění struktury je náročné a nákladné
- Znalost struktury transmembránových proteinů je důležitá např. při vývoji nových léků
- Je však užitečné vědět alespoň základní topologii proteinu:
  - Části, které jsou **uvnitř buňky** (*cytoplasmatic*)
  - Části, které jsou **uvnitř membrány** (*transmembrane*)
  - Části, které jsou **vně buňky** (*extracellular*)



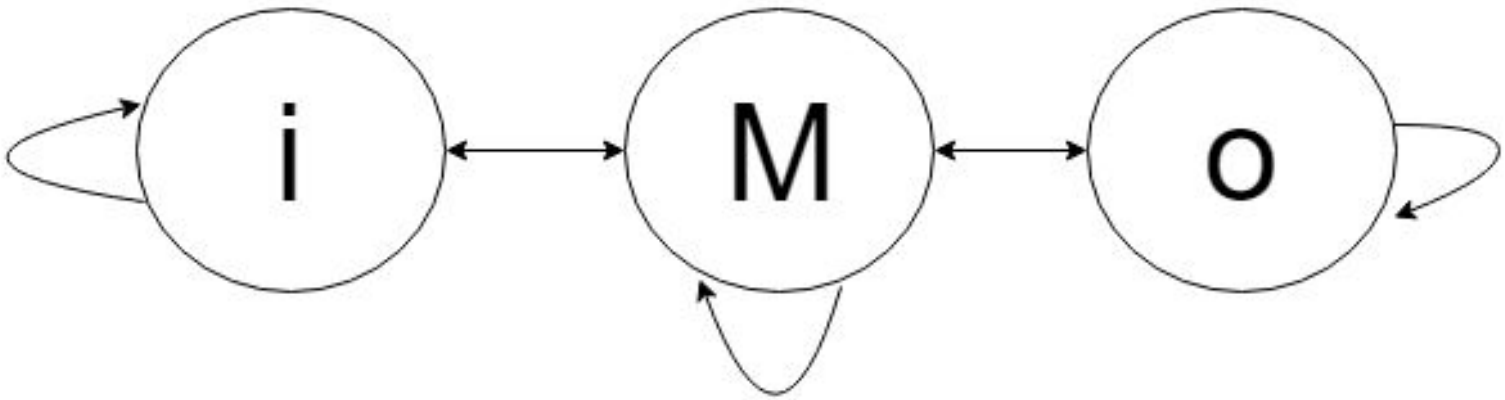
# State-of-the art metoda TMHMM

- Založen na skrytých markovských řetězcích
- Velká přesnost detekce
  - *July 2001: TMHMM has been rated best in an independent comparison of programs for prediction of TM helices*
- Pro akademické účely zdarma ke stažení
- Funguje i jako webová služba
- Dataset 160 proteinů
  - Řetězece aminokyselin jsou anotován:
    - i - uvnitř buňky
    - M - transmembránová šroubovice
    - o - vně buňky
- Doporučeno známým doktorem z Fyziologického ústavu AV ČR

# Učení HMM s učitelem

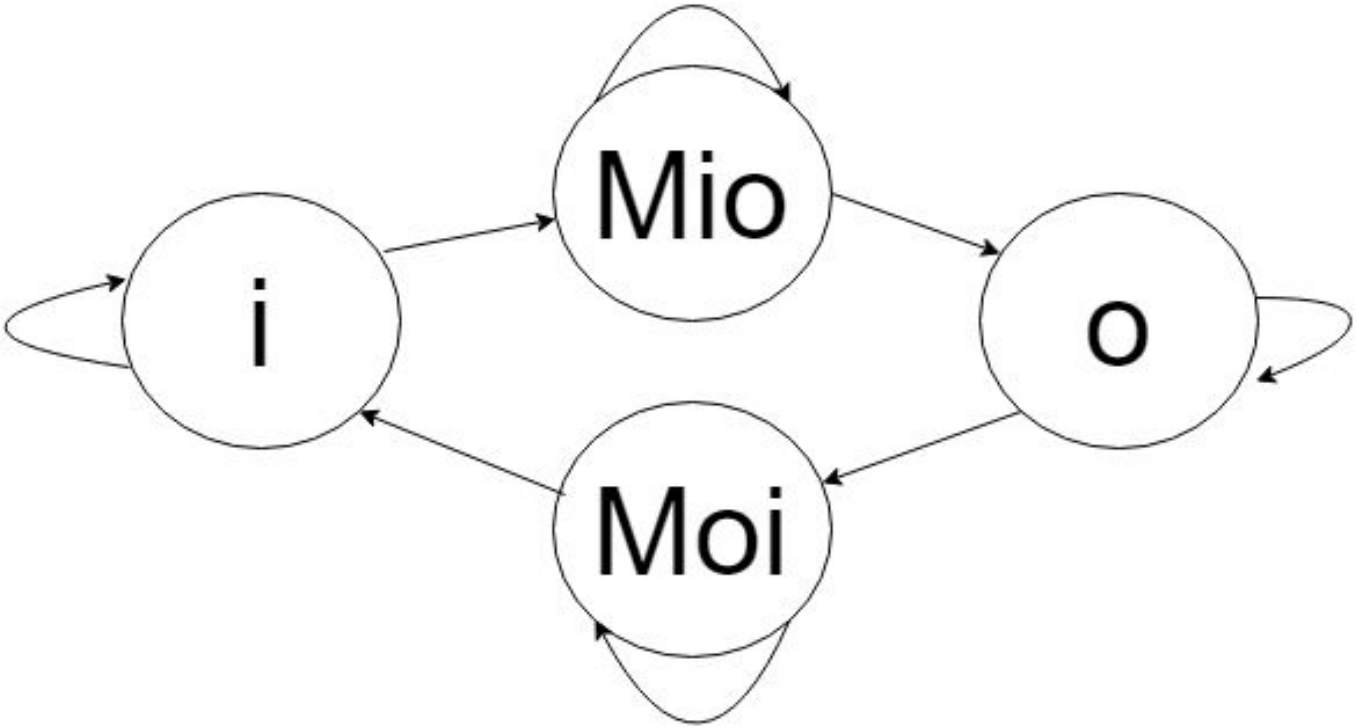
- Nejpravděpodobnější průchod stavy modelu na základě pozorovaných hodnot (aminokyselin)
- Dataset:
  - 160 proteinů různé velikosti s anotacemi
  - 10-fold cross validation
- Python
  - Knihovny hmmlearn ani pomegranate nebyly vhodné
  - Proto jsem využil zdrojového kódu pro HMM z MIT
  - http://www.mit.edu/course/6/6.863/OldFiles/python/old/nltk-contrib-1.4.2/build/lib/nltk_contrib/unimelb/tacohn/hmm.py
- 3 modely:
  - Naivní model
  - Naivní model 2
  - Pokročilý model

# Naivní model




- 3 stavy
- 20 emitovaných symbolů (aminokyseliny)
  - ACDEFGHIKLMNPQRSTVWY

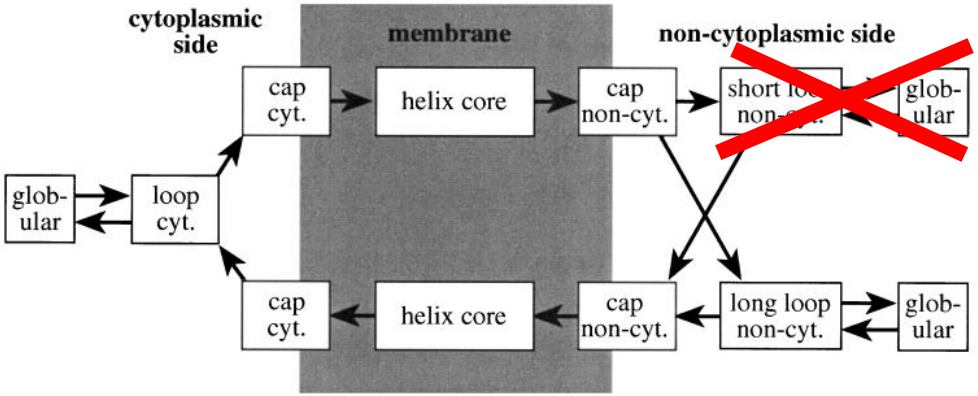
# Naivní model 2

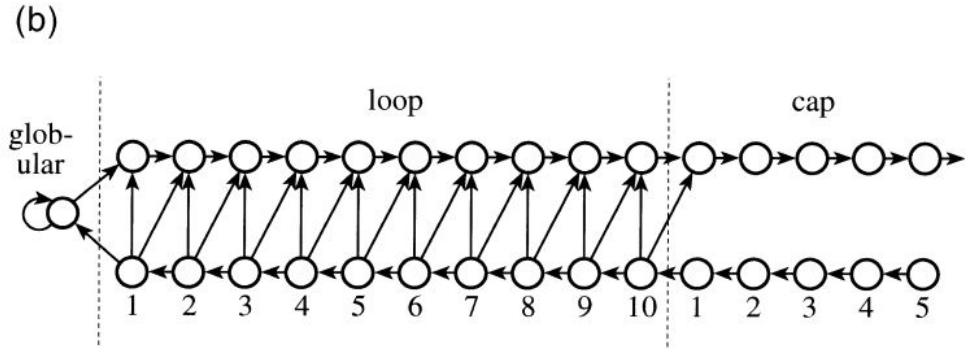


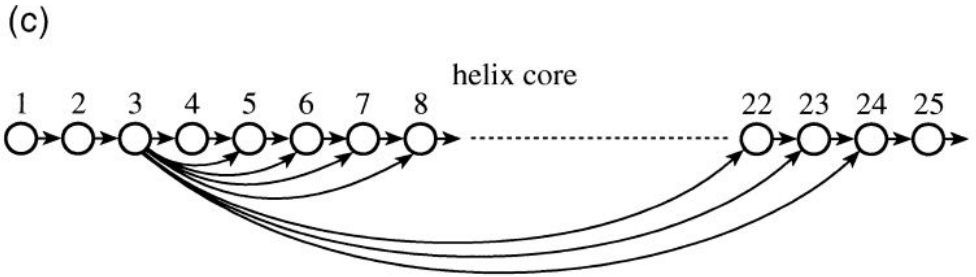

- 4 stavy
  - Zahrnuje v sobě znalost toho, že nemohou být dvě sekvence “i” oddělené jednou sekvencí “M”
  - Stavy “Mio” a “Moi” odpovídají značce “M” ve výstupu
- 20 emitovaných symbolů (aminokyseliny)
  - ACDEFGHIKLMNPQRSTVWY

# Pokročilý model

- Pokus implementace modelu z článku
- Velké množství možných stavů (> 100)
- Je třeba mapovat sekvence “iii...iiiMM...MMoo…” na sekvenci stavů a zpět



# Výsledky experimentů

| | Naive Model | | Naive Model v2 | | Advanced Model | | TMHMM | |
|---|---|---|---|---|---|---|---|---|
| success rate: | 72.25% | | 73.88% | | 85.64% | | | |
| total helixes | 696 | | 696 | | 696 | | 696 | |
| over predicted | 9 | 1.29% | 17 | 2.44% | 34 | 4.89% | 17 | 2.44% |
| under predicted | 150 | 21.55% | 142 | 20.40% | 27 | 3.88% | 19 | 2.73% |

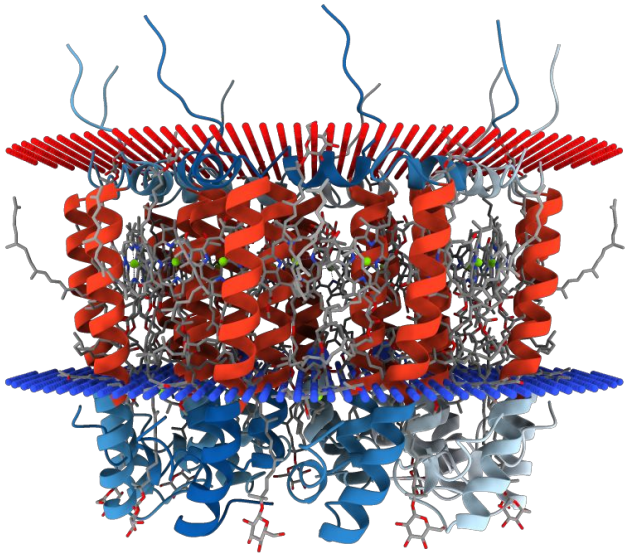
- Naivní modely fungovaly překvapivě dobře
- Pokročilý model je výrazně lepší
  - Rádově nižší počet nedetekovaných šroubovic
- TMHMM využívá specifičtějšího návrhu stavů
  - Stále však není schopen určit přesné hranice membrány

# Příklad: P26789 (LHA4_RHOAC)

```
input: MNQGKIWTVVNPAIGIPALLGSVTVIAILVHLAILSHTTWFPAYWQGGVKKAA
label: iiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMMooooooooooooooooo
pred.: iiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMMooooooooooooooooo
TMHMM: iiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMMooooooooooooooooo

error:        5.66 %

>     total helixes:  1
>     overpredicted:  0 (0.0 %)
>     underpredicted: 0 (0.0 %)
```



Zdroj: http://www.rcsb.org/pdb/explore/explore.do?structureId=1NKZ

# Možná zlepšení

- Při učení nebyly použity žádné nemembránové proteiny
  - Model není vhodný pro použití při klasifikaci membránových a nemembránových proteinů
- Ani pokročilý model neodpovídá přesně modelu TMHMM
  - Došlo k několika zjednodušením modelu
  - Implementováním pokročilého modelu by mohlo dojít k dalšímu zpřesnění predikcí

# Zdroje


- A. Krogh, B. Larsson, G. von Heijne, and E. L. L. Sonnhammer.
  *Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes.*
  Journal of Molecular Biology, 305(3):567-580, January 2001.
  Available online: http://www.cbs.dtu.dk:80/~krogh/TMHMM/
- De Fonzo, Valeria, Filippo Aluffi-Pentini, and Valerio Parisi.
  *Hidden Markov models in bioinformatics.*
  Current Bioinformatics 2.1 (2007): 49-61
  Available online:
  https://pdfs.semanticscholar.org/82ee/11e75ce2a0fae98d26cc1d41d9a47c41d4fc.pdf
- http://gpcrdb.org
- http://www.rcsb.org