



Bioinformatika és genomanalízis az orvostudományban

Bioinformatikai modellek

Cserző Miklós

2020

<https://semmelweis.zoom.us/j/96102872458?pwd=Rk1PL2tqS21sdIUwc3B4eDFCZkNKQT09>

A mai előadás

- A predikció jelentősége a biológiában
- Egyszerű statisztikai modellek
 - Kyte-Doolittle hidrofóbicitás skála
 - Chou–Fasman szerkezet predikció
- Nem-annyira-egyszerű modellek
 - Rejtett Markov modell
 - Neurális háló
 - SVM modell



Predikció

- Van sok adatunk
- Feltételezünk valamilyen szabályszerűséget a rendszerben
- Következtetünk ezekre
- Modellt építünk ezek alapján
- Kipróbáljuk a modellt
- Olyan rendszerek viselkedését jósoljuk meg, amelyekre nincs közvetlen adatunk

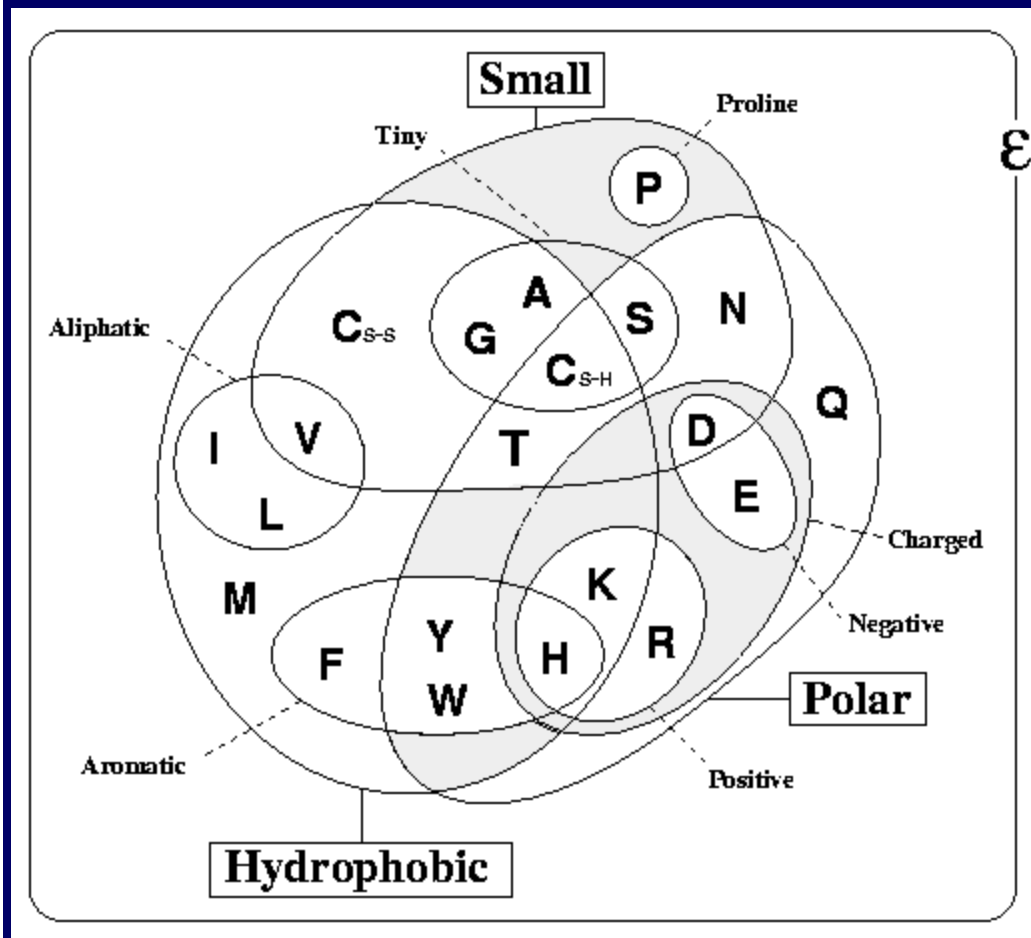
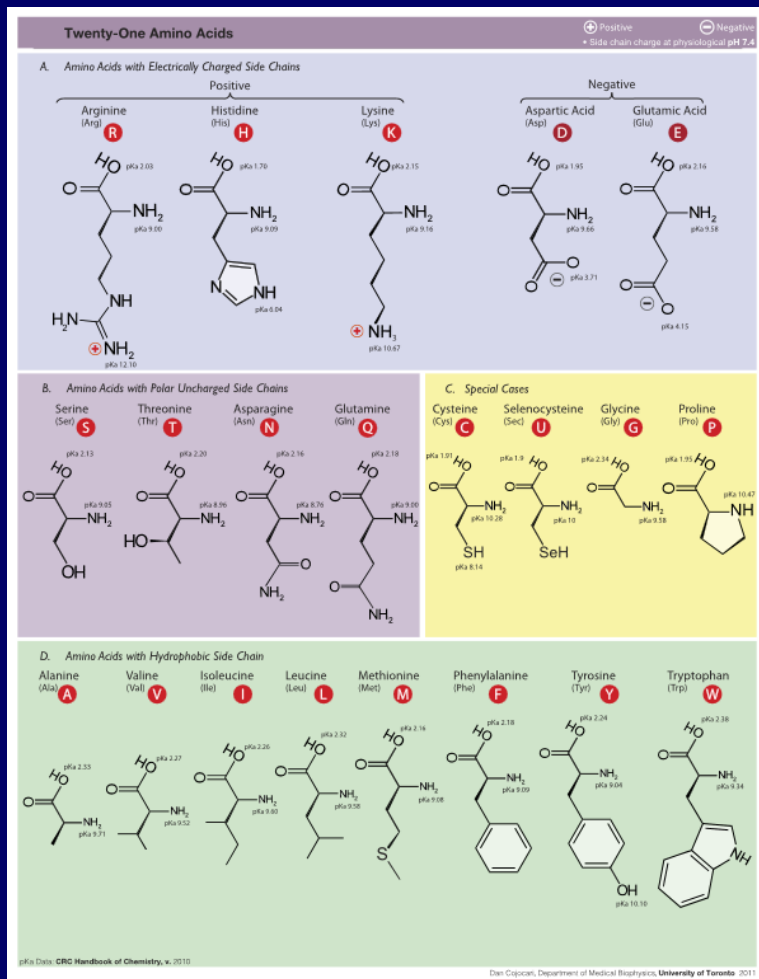
A hidrofóbicitás

- A fehérje folding hajtóereje
- A sejten belüli elhelyezkedés szempontjából kitüntetett – membrán fehérjék
- A sejt és környezete közti jelátvitelben kitüntetett szerep
- Ezen keresztül nagy gyógyászati jelentőséggel bír

Kyte-Doolittle modell

- Az egyes aminosavak hidrofóbicitása adott
- Ez egy additív mennyiség
- Ha a szekvenciában sok hidrofób aminosav van egy helyen koncentráva az egész fragmens hidrofób lesz
- A hidrofób részek az első számú jelöltek a transzmembrán szakaszok azonosítására

Aminósavak tulajdonságai



GenomeNet

Database: AAindex

Entry: KYTJ820101

LinkDB: KYTJ820101

H KYTJ820101

D Hydropathy index (Kyte-Doolittle, 1982)

R LIT:0807099 PMID:7108955

A Kyte, J. and Doolittle, R.F.

T A simple method for displaying the hydropathic character of a protein

J J. Mol. Biol. 157, 105-132 (1982)

C JURD980101	0.996	CHOC760103	0.964	OLSK800101	0.942
JANJ780102	0.922	NADH010102	0.920	NADH010101	0.918
DESM900102	0.898	EISD860103	0.897	CHOC760104	0.889
NADH010103	0.885	WOLR810101	0.885	RADA880101	0.884
MANP780101	0.881	EISD840101	0.878	PONP800103	0.870
WOLR790101	0.869	NAKH920108	0.868	JANJ790101	0.867
JANJ790102	0.866	BASU050103	0.863	PONP800102	0.861
MEIH800103	0.856	NADH010104	0.856	PONP800101	0.851
PONP800108	0.850	CORJ870101	0.848	WARP780101	0.845
COWR900101	0.845	PONP930101	0.844	RADA880108	0.842
ROSG850102	0.841	DESM900101	0.837	BLAS910101	0.836
BIOV880101	0.829	RADA880107	0.828	BASU050101	0.826
KANM800104	0.824	LIFS790102	0.824	CIDH920104	0.824
MIYS850101	0.821	RADA880104	0.819	NAKH900111	0.817
CORJ870104	0.812	NISK800101	0.812	FAUJ830101	0.811
ROSM880105	0.806	ARGP820103	0.806	CORJ870103	0.806
NADH010105	0.804	NAKH920105	0.803	ARGP820102	0.803
CORJ870107	0.801	MIYS990104	-0.800	CORJ870108	-0.802
KRIW790101	-0.805	MIYS990105	-0.818	MIYS990103	-0.833
CHOC760102	-0.838	MIYS990101	-0.840	MIYS990102	-0.840
MONM990101	-0.842	GUYH850101	-0.843	FASG890101	-0.844
RACS770102	-0.844	ROSM880101	-0.845	JANJ780103	-0.845
ENGD860101	-0.850	PRAM900101	-0.850	JANJ780101	-0.852
GRAR740102	-0.859	PUNT030102	-0.862	GUYH850104	-0.869
MEIH800102	-0.871	PUNT030101	-0.872	ROSM880102	-0.878
KUHL950101	-0.883	GUYH850105	-0.883	OCEM770101	-0.899

I	A/L	R/K	N/M	D/F	C/P	Q/S	E/T	G/W	H/Y	I/V
	1.8	-4.5	-3.5	-3.5	2.5	-3.5	-3.5	-0.4	-3.2	4.5
	3.8	-3.9	1.9	2.8	-1.6	-0.8	-0.7	-0.9	-1.3	4.2

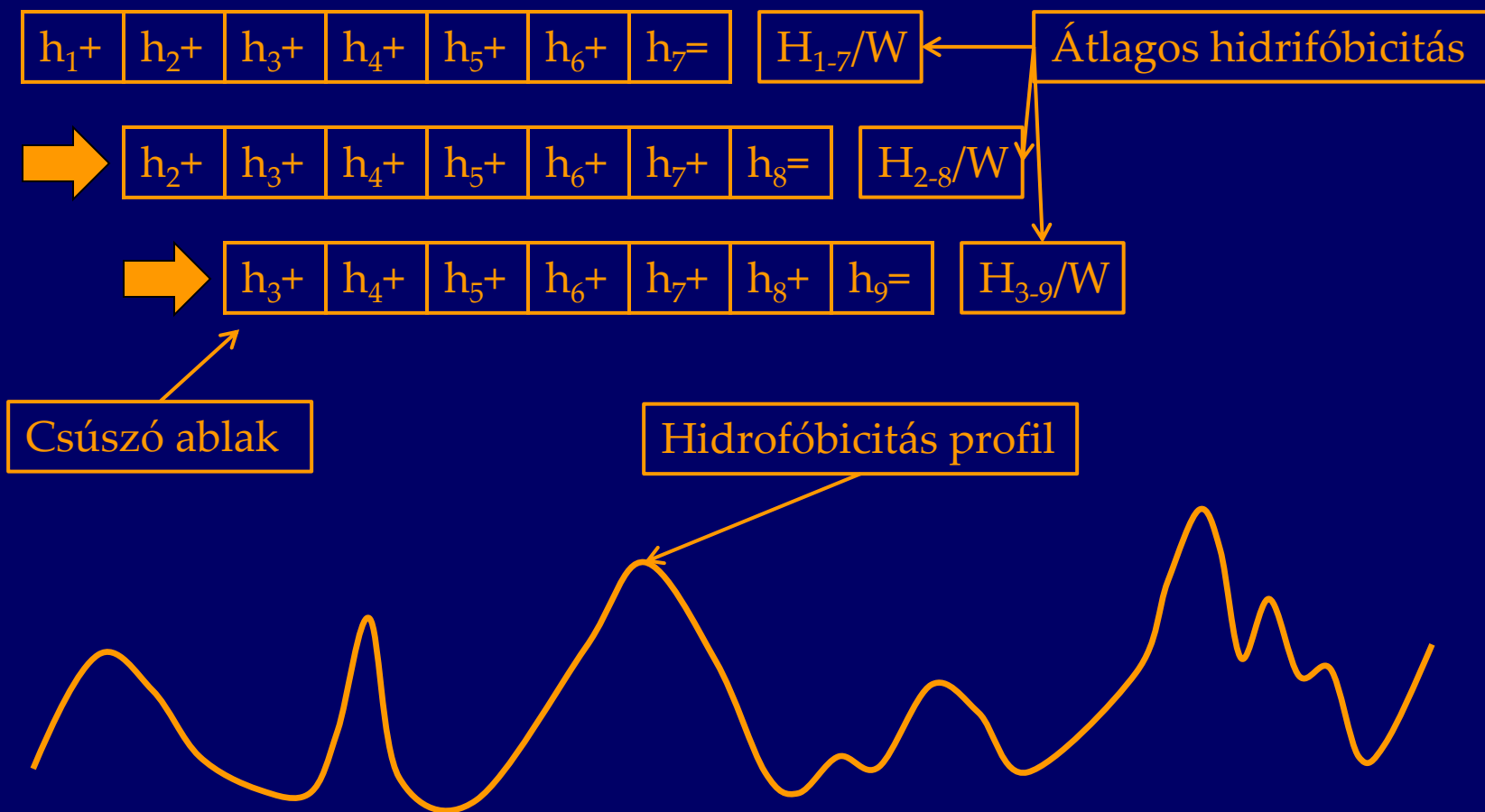
//

DBGET integrated database retrieval system

A hidrofóbicitásról

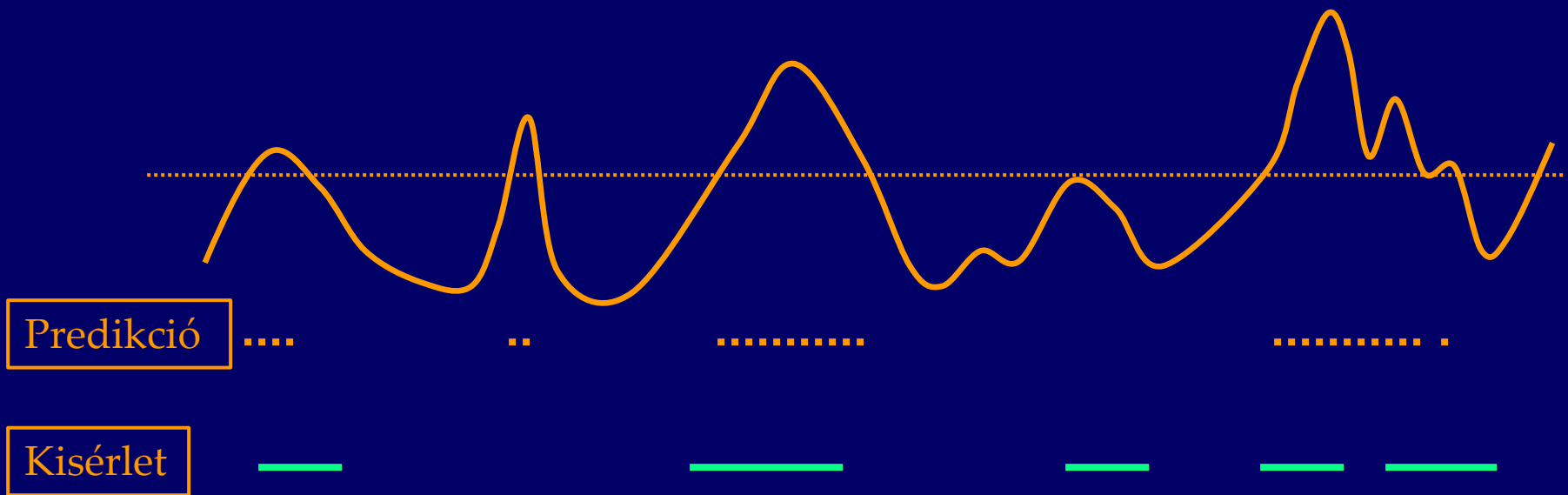
- Nagy számú és sokfajta mikroszkópikus erő makroszkópikus eredője
- Meg lehet mérni fizikai-kémiai módszerekkel
- Többféle képpen is
- Pl. megoszlási hányadoson keresztül
 - Mi legyen Az oldószer?
 - Mi legyen a minta?

A számítás menete

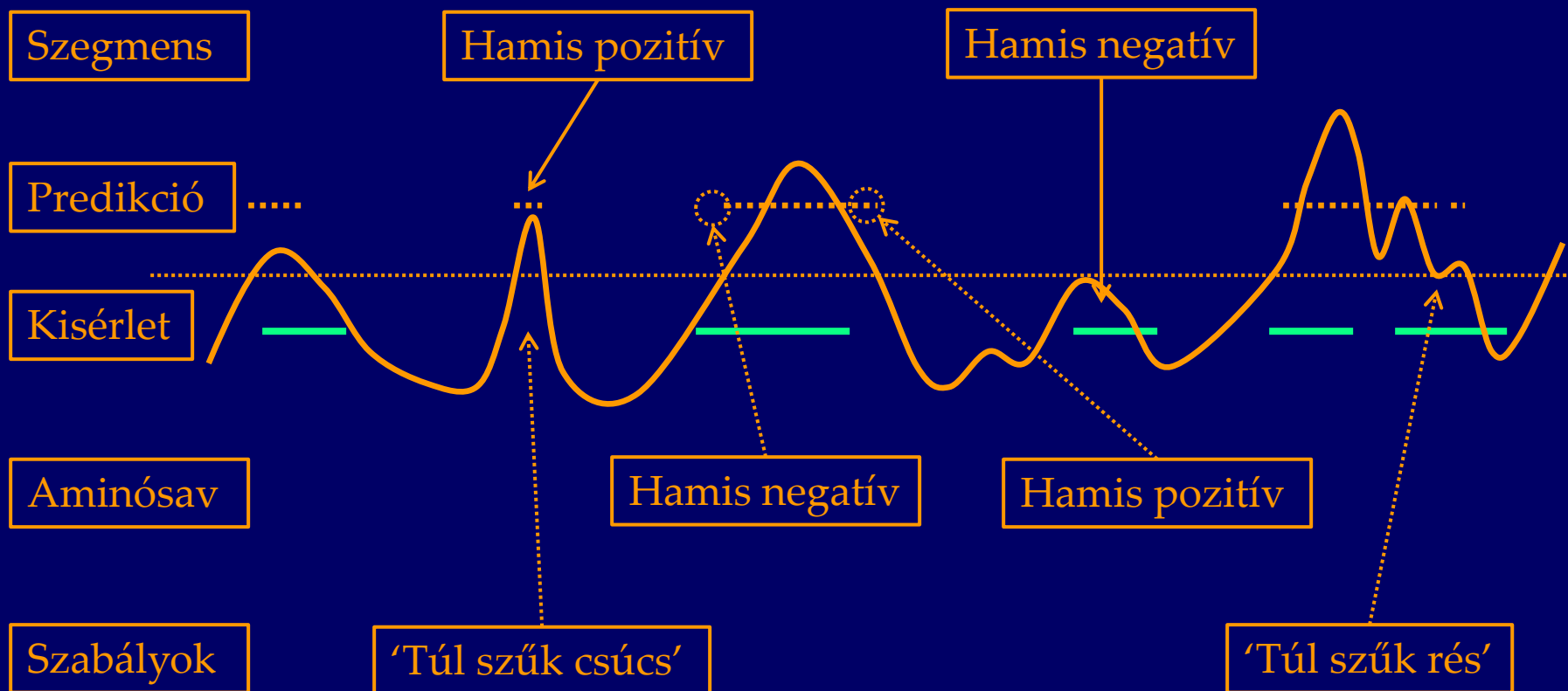


Kiértékelés

Paraméterek: skála, ablak, levágás



Korrektíós lehetőségek

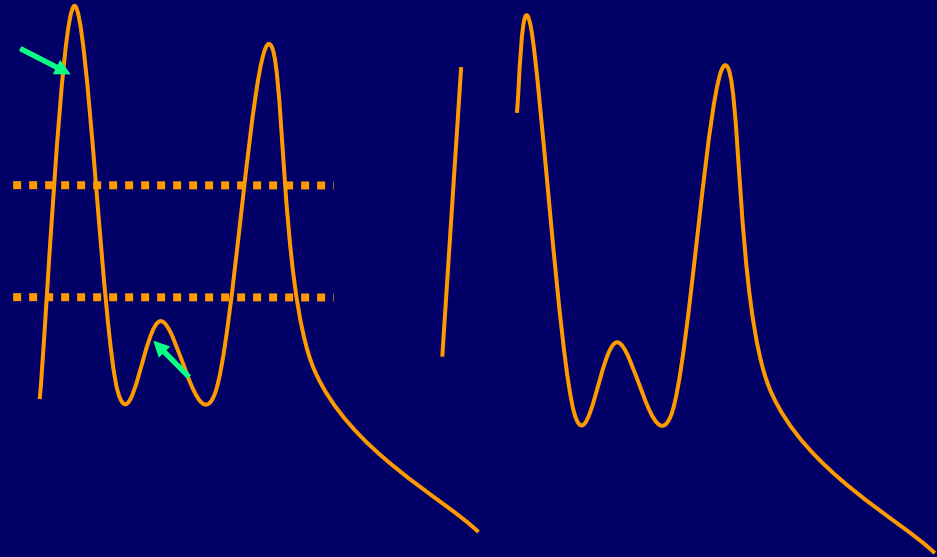


Korrektációs hibák

- A hibaforrások:
 - Túl kevés kísérletes adat
 - Pontatlan kísérletes adatok
- Transzmembrán fehérjék esetén mindkét problémával találkozunk
 - TM fehérjék érzékenyek
 - Nehéz kísérletezni velük
 - Így kevés és pontatlan adatunk van

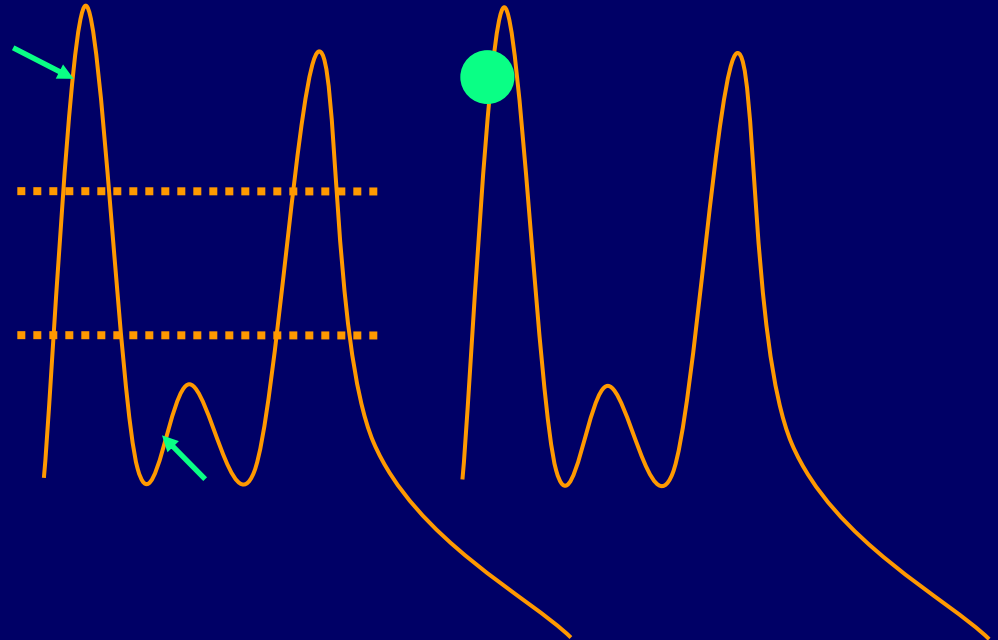
Limitált proteolízis

- Specifikus proteázok
- Csak bizonyos szekvenmciáknál hasítanak
- Ilyenek vagy vannak a fehérjében vagy nem
- Vagy elérhetők vagy nem



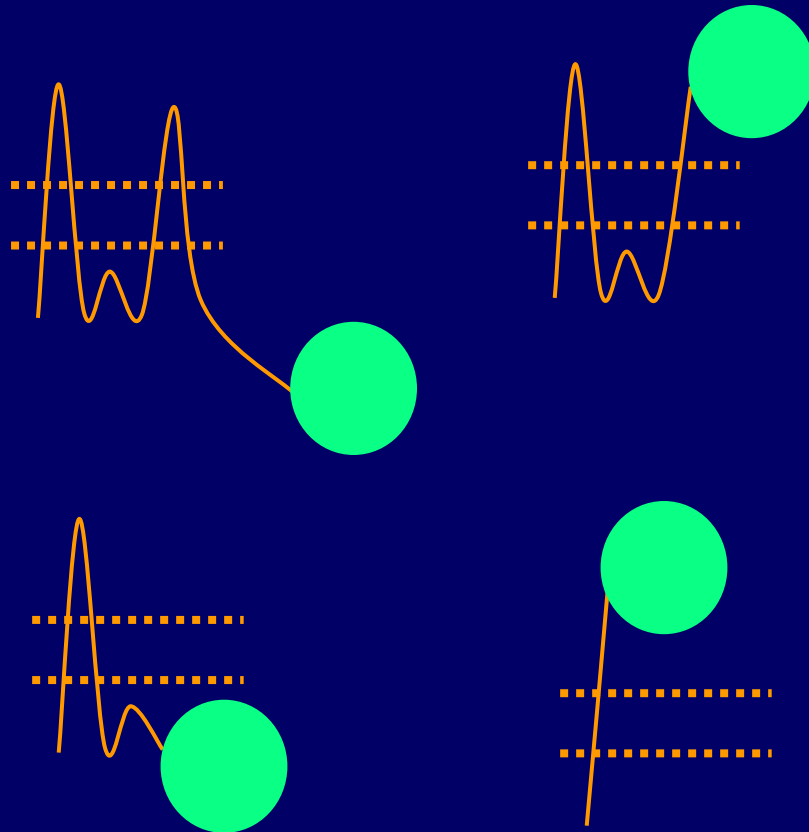
Jelölés

- Specifikus festékek
- Bizonyos aminosavakat megjelölnék kovalens kötésekkel
- Ilyenek vagy vannak a fehérjében vagy nem
- Vagy elérhetőek vagy nem



Fúziós fehérje

- A fehérjét megtoldjuk egy riporter fehérjével
- Ez egy specifikus reagenssel jelet ad
- Az eredeti fehérjéből egyre nagyobb darabokat vágunk le a toldás előtt
- Ha a reagens és a riporter a membrán azonos oldalán van – jelet kapunk



A közös probléma

- Nagyon körülményes a kísérlet
- Csak kivételes esetben működik
- Nem ad pontos eredményt
- Inkább csak egy predikció eredményével ér fel

A Chou–Fasman predikció

- A fehérjék másodlagos szerkezetét próbálja megjósolni
- Négy állású modellt alkalmaz: α -helix, β -redő, γ -hurok és a maradék
- Megoldott fehérje szerkezetek alapján számolt 'szerkezet képző hajlamot' használ
- Az elv hasonlít a K-D TM predikcióra, csak négy profil alapján készít predikciót

A modell

- 'x' aminosav α -helix képző 'hajlama': 'x' gyakorisága helixben v.s. nem-helixben
- C-F paraméterek: $p(x)_\alpha = f(x)_\alpha / f(x)_\epsilon$
- A többi szerkezeti elemre analóg módon számolt paraméter
- Ezek alapján predikciós profilt készítünk
- A csúcsok jelentik a szerkezeti elemek magjait
- Ezeket addig szélesítjük, míg beleütközünk a következő szerkezeti elembe

GenomeNet

Database: AAindex

Entry: CHOP780201

LinkDB: [CHOP780201](#)

H CHOP780201

D Normalized frequency of alpha-helix (Chou-Fasman, 1978b)

R PMID:[364941](#)

A Chou, P.Y. and Fasman, G.D.

T Prediction of the secondary structure of proteins from their amino acid sequence

J Adv. Enzymol. 47, 45-148 (1978)

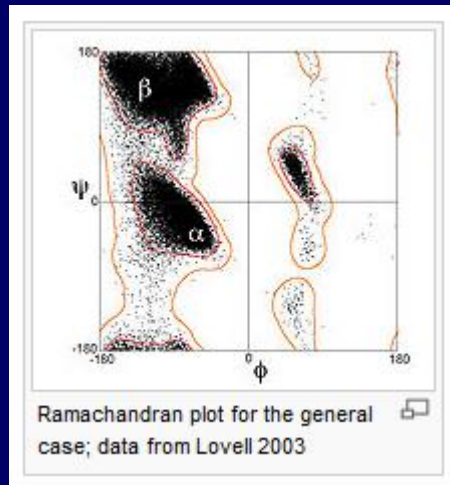
C	PALJ810102	0.981	ROBB760101	0.969	ISOY800101	0.959				
	MAXF760101	0.956	KANM800101	0.956	TANS770101	0.947				
	BURA740101	0.917	GEIM800101	0.912	KANM800103	0.912				
	NAGK730101	0.886	LEVW780104	0.886	PALJ810101	0.881				
	QIAN880106	0.874	FRAM900102	0.873	LEVW780101	0.873				
	GEIM800104	0.868	RACS820108	0.868	AURR980108	0.867				
	AURR980109	0.859	AURR980112	0.856	CRAJ730101	0.851				
	QIAN880107	0.843	BEGF750101	0.841	QIAN880105	0.835				
	AURR980114	0.828	AURR980115	0.816	AURR980110	0.814				
	PALJ810109	0.814	AURR980111	0.813	ROBB760103	0.806				
	MUNV940101	-0.802	CRAJ730103	-0.808	ROBB760113	-0.811				
	MUNV940102	-0.812	CHAM830101	-0.828	NAGK730103	-0.837				
I	A/L	R/K	N/M	D/F	C/P	Q/S	E/T	G/W	H/Y	I/V
	1.42	0.98	0.67	1.01	0.70	1.11	1.51	0.57	1.00	1.08
	1.21	1.16	1.45	1.13	0.57	0.77	0.83	1.08	0.69	1.06

//

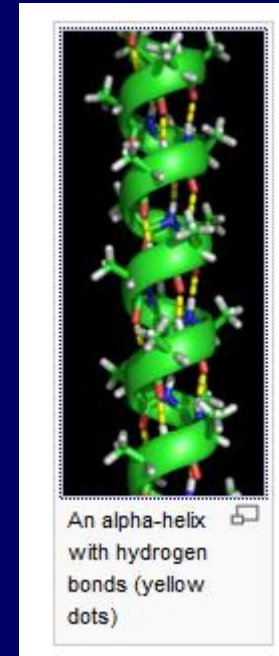
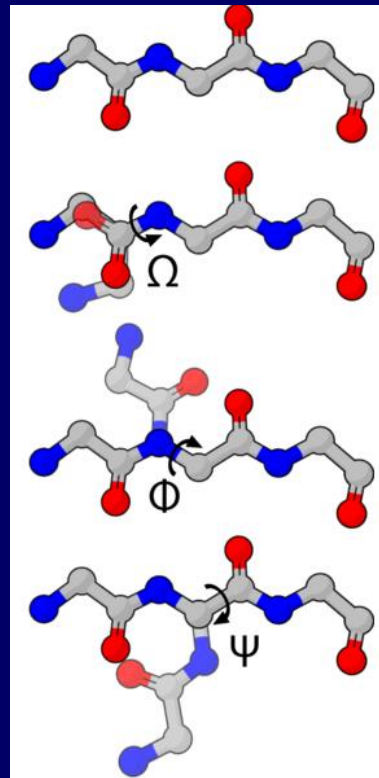
DBGET integrated database retrieval system

A modell kalibrálása – lehetséges definíciók

Ramachandran plot



Szerkezeti definíció





Problémák

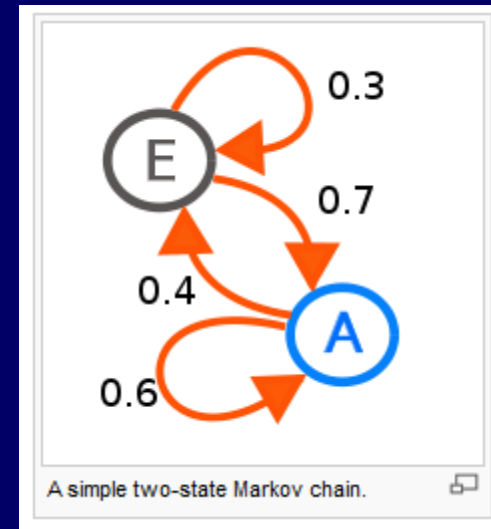
- Már a szerkezeti elemek azonosítása sem egyértelmű
- Az adatbázis nagyon kicsi volt az eredeti módszer kidolgozásakor – kb. 3500 aminosav
- A modell felbontása alacsony
- Figyelmen kívül hagy további szerkezeti elemeket

Tanulság

- A K-D és C-F módszer nagyon egyszerű, de pontatlan
- A modell paramétereit optimalizálni kell kísérletes adatok felhasználásával
- Kerülni kell a modell 'túlillesztését'
- Pontatlan adatokra alapozott rendszer szükségszerűen pontatlan
- Ha egy módszer egyszerű mindenki használni fogja, akkor is, ha megbízhatatlan
- Ne használjuk ezeket a predikciókat!!!

Markov lánc, Markov modell

- Van egy rendszerünk
- A rendszer véges számú állapotot vehet fel – állapottér
- A rendszer az állapottér egyik pontjából a másikba adott valószínűséggel jut át



AAEEAAEEAAAAEEAAEAAAAAEEE
AAEEAAEAEAEAAEAAAEAAEAAA

Milyen az idő ma?

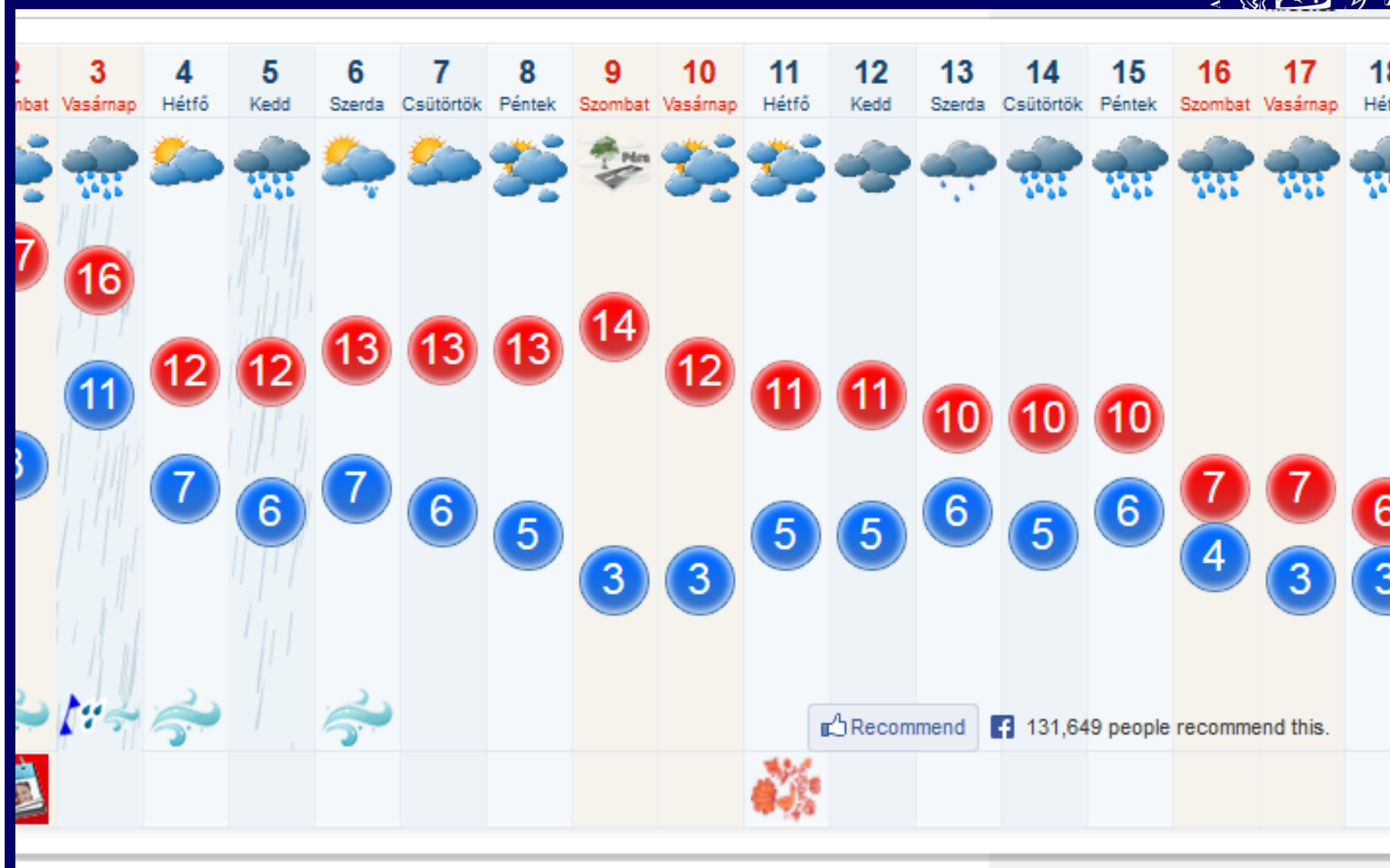
➤ Az időjárás paramétereit:

- hőmérséklet
- légnyomás
- páratartalom
- szélirány
- szélerősség



És holnap?

- Ugyan olyan, mint ma
 - 60%-ban helyes jóslat
- Melegebb vagy hidegebb?
 - Egy kicsivel – lehetséges
 - Sokkal – talán, de ez ritkább
- A többi paraméterre is hasonló megfontolást alkalmazunk



Mi van a befektetésekkel?

- A tőzsdén forognak az értékpapírok
- A papírok értéke változik
 - Az egyik drágul, a másik veszít az értékéből
- Ha meg tudom jósolni a jövőbeli árfolyamokat, akkor jól járok
- A papírok értéke rövid távon inkább kicsit változik, a nagy változás ritka
- Elégé hasonlít az időjárásra....



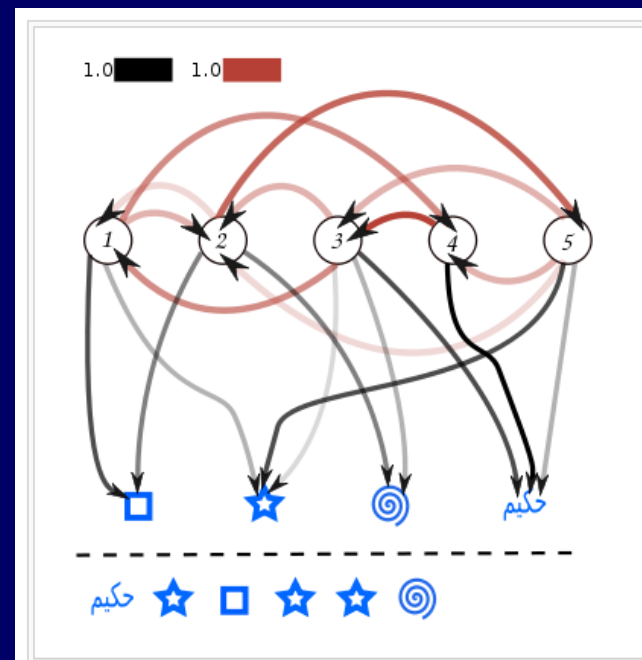
Adatok forrása: **Portfolio**
Financial Markers

Lehetséges modell

- Mindkét folyamatban vannak közös staitisztikus elemek
 - A jelen állapot ismert
 - A folytonos paramétereket kerekíteni kell
 - Az adott állapot megváltozása
 - kismértékben - valószínű
 - nagymértékben - valószínűtlen
- A 'ritán' és a 'soha' két különböző dolog!!!
 - Rövid távra könnyebb jósolni
 - Ne tőzsdézzünk és hordjunk magunkkal esernyőt!

Rejtett Markov modell

- Markov modell egy rejtett réteggel megtoldva
- Az állapotokat nem érzékeljük közvetlenül
- Csak az eredményt látjuk



Javított időjárás modell

- Hiányoznak további adatok:
 - Magaslégkör állapota
 - Tengeráramlások adatai
- Ezeket nem tudjuk mérni
- Kell még az évszak is
- A tapasztalt helyi időjárás ezek következménye
- A rejtett Markov modell jobb lehet

HMM alapú bioinformatika

- A Markov lánc eredetileg a rendszer időbeli viselkedését írja le
- A biológiában a szekvencia az érdekes – fehérje vagy nukleinsav
- A szekvencia egy nem ismert tulajdonságát modellezzük, aminek egy megnyilvánulása az adott szekvencia
- Pl. <http://www.enzim.hu/hmmtop/>

[Home](#) | [Documentation](#) | [Help](#) | [Download](#) | [Advanced](#) | [Submit](#)

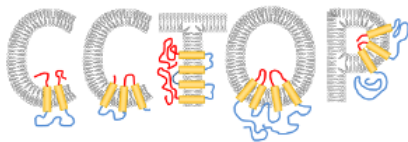
Copyright © G. E. Tusnády, 2001

HMMTOP

*Prediction of transmembrane helices
and topology of proteins*

Version 2.0

by G. E. Tusnády



Constrained Consensus TOPology prediction server



Documents



Predictions

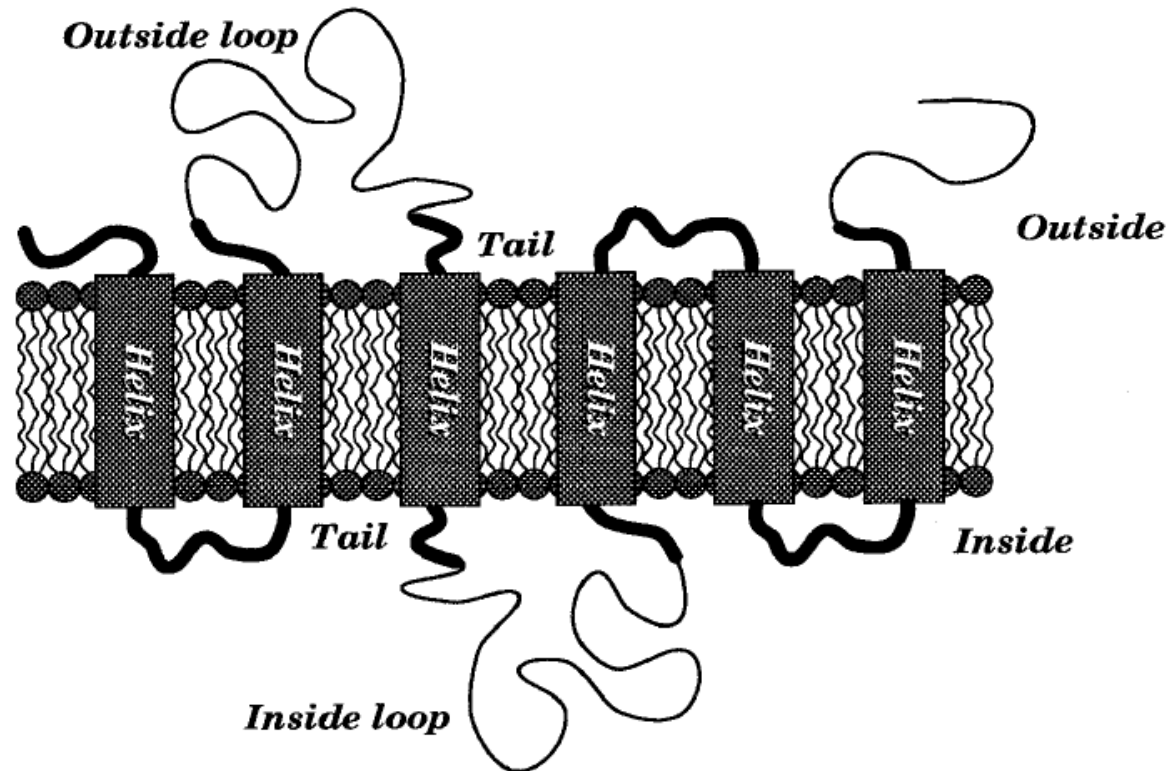
Visitors: 1442745
Prediction count: 279760
Server version: s.1.00
Last changed: 13.07.2017

Welcome to the Constrained Consensus TOPology prediction server.

Given the amino acid sequence of a putative α -helical transmembrane protein, CCTOP predicts its topology i.e. localization of membrane spanning regions and orientation of segments between them. The prediction is a consensus of 10 different methods enhanced with available structural and experimental information of any homologous proteins in the **TOPDB** database. CCTOP was tested on a benchmark set containing 170 proteins with known structure and achieved the highest accuracy among state-of-art and consensus methods.

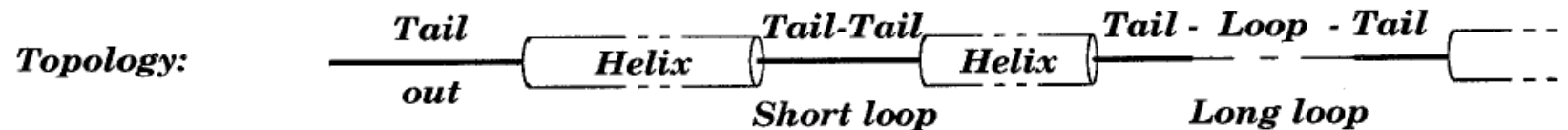
You can find more information how CCTOP works in the **description** or in the **article**. A short description about the server can be found in the **manual**.

You can submit protein sequences **here**.



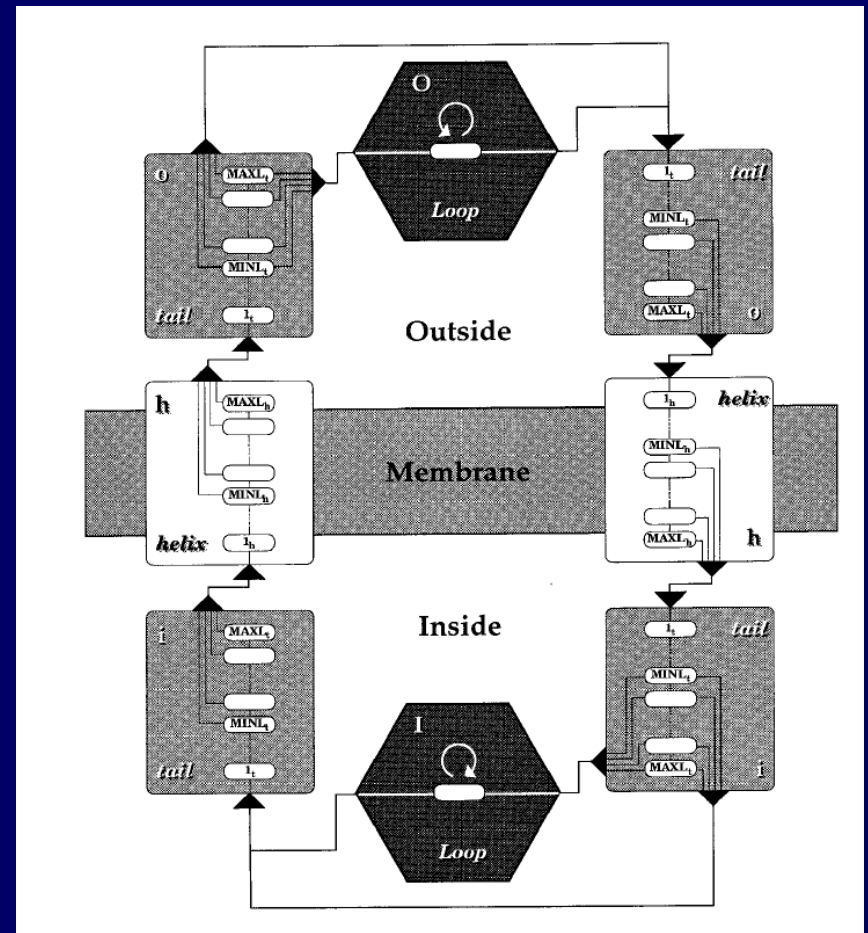
Amino acid seq: MGDVCDTEFGILVA...SVALRPRKHGRWIV...FWVDNGTEQ...PEHMTKLHMM...

State seq: oooooooooohhhhh...hhhhiiiiiiihhh...hhhooooOO...OOOooooohhh...



A modell

- A modell 5 állapotot különböztet meg
- Tartalmaz terjedelmi korlátot is
- A szerver több mint 20 éve üzemelt
- Ma egy konszenzus szerver fut helyette

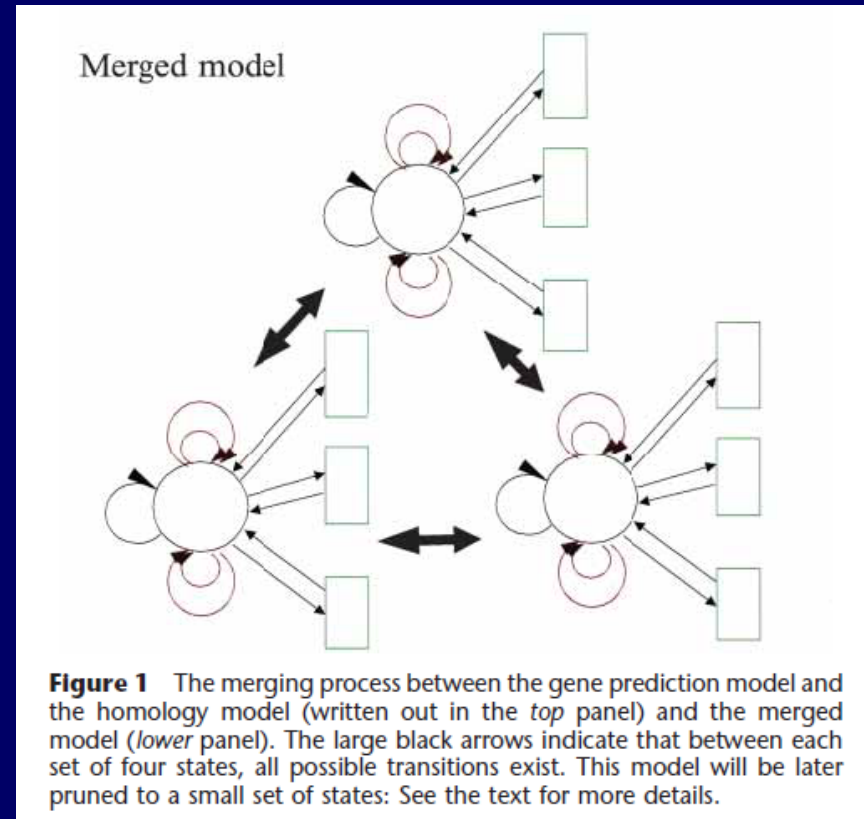


A teljes leírás

- "G.E Tusnády and I. Simon (1998) Principles Governing Amino Acid Composition of Integral Membrane Proteins: Applications to Topology Prediction." *J. Mol. Biol.* **283**, 489-506

GeneWise – páros HMM

- Kell egy modell a
 - Az exon/intron szerkezet becslésére
 - DNS – fehérje fordításra
 - A fehérje – lefordított fehérje összehasonlításra



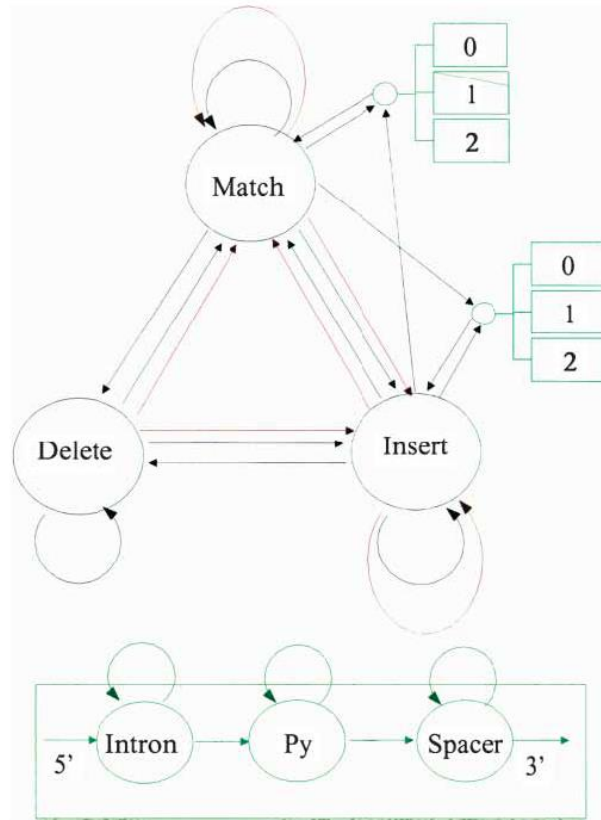


Figure 2 The GeneWise21:93 algorithm. The circles represent states, and the arrows between them transitions. Black transitions are standard protein transitions, red transitions are frameshifting transitions, and green transitions are intronic transitions. Introns are each built of three states, listed at the bottom of the figure.



Referencia

- GeneWise and Genomewise, Ewan Birney, Michele Clamp and Richard Durbin, *Genome Res.* 2004 14: 988-995

A modellek paramétereai

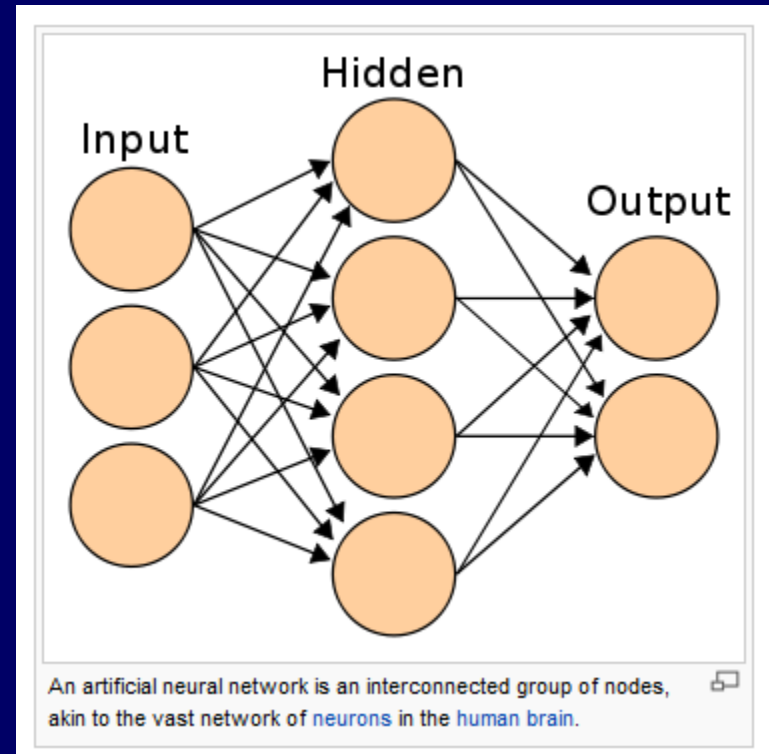
- A HMM nem használható a belső paramétereai nélkül
 - Állapot-átmenet valószínűség
 - Kimeneti valószínűség
- Ezeket meg kell tanítani a modellnek
- Nincs általános megoldás, csak közelítő
- A modellt 'visszafelé' kell működtetni
- Ehhez kell a 'tanuló' készlet
 - Ilyent már láttunk – K-D és C-F módszer

Sok HMM alapú módszer van

- Szekvencia motívum keresés
- Génszerkezet predikció
- Fehérje szerkezet predikció
- Többszörös szekvencia-illesztés
- Szerkezeti motívum keresés
- És még sok más.....

Neurális hálózat

- Biológiai eredetű statisztikai modell
- A modell egységei rétegekben állnak
- Az egység az előző rétegből kapja a bemenetét
- És a következő réteg egységeinek adja tovább
- A csatolás súlyozott



Az elrendezés

- A rétegek száma lehet több is
- A rétegek mérete eltérhet
- A modellt tanítani kell
- Hasonló feladatokat old meg, mint a HMM
- A 90-s években igen népszerű volt
- Mostanában kevesebbet használják

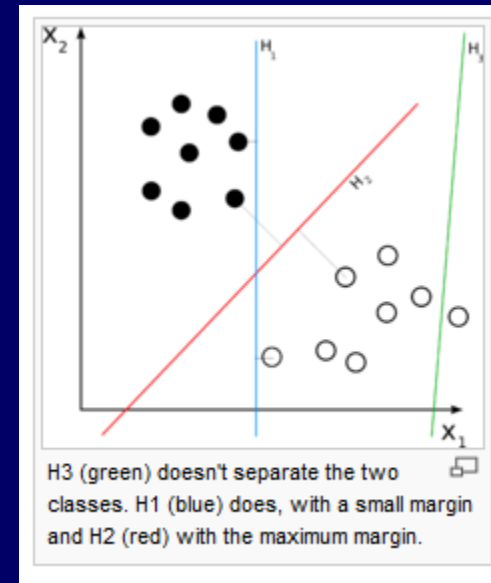
Javított időjárás előrejelzés

- Tudjuk, milyen az idő Budapesten ma
- Tudjuk a mai időt
 - Bécsben
 - Krakkóban
 - Kievben
 - Belgrádban
- Milyen lesz az idő holnap Budapesten?



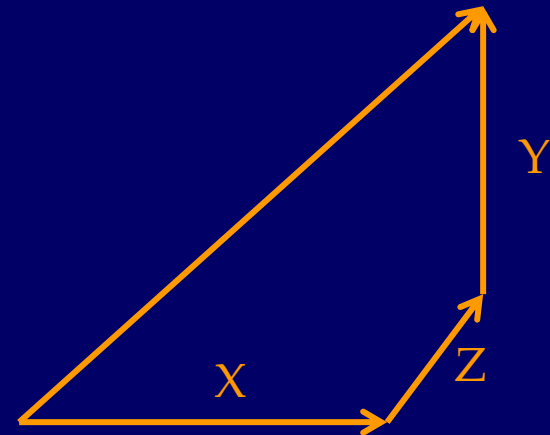
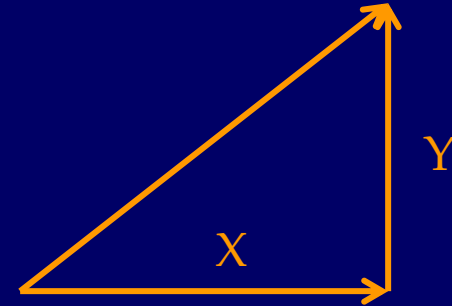
Support Vector Machine

- Van két pontthalmazunk a síkon
- Keressük azt az egyenest, amely elválasztja egymástól a két halmazt
- És a lehető legszélesebb elválasztósávot hagy



Vektorgeometriai megközelítés

- Téből van
 - 1D – egy koordináta
 - 2D – x, y koordináta
 - 3D – x, y, z
 - Sőt több...
 - N koordináta – vektor
- A vektorgeometria egy régi, jól ismert és egyszerű tudomány



SVM a biológiában

- A modell használatához kell egy tanító adatsor
 - Egy pontsor 'A' tulajdonsággal
 - Egy pontsor 'B' tulajdonsággal
- Ez alapján kalibráljuk a modellt
- A további pontokról el tudjuk dönteni, hogy az 'A' vagy 'B' halmazba tartoznak-e

Megjegyzések

- A modell eredeti formájában csak két halmaz felismerésére alkalmas
- A biológiában szekvenciáink vannak és nem vektoraink
 - A szekvenciát át kell alakítani vektorokká
 - Meg kell találni a monomerek számszerűsíthető tulajdonságait
- A modell igen népszerű mostanában

SVM alapú szerkezet predikció

- C-F szerű módszer SVM alapon
- A modell felbontása: helix / nem helix
- Honnan veszünk számszerű adatokat?
 - AAindex adatbázis
- Egyes aminosavakat nem elég vizsgálni
 - 'X' aminosav vektora mindig egy pontba mutat, akár helixben van, akár nem
- Figyelembe kell venni a környezetet is
 - 'X' aminosav vektora más környezetben más lesz

A lehetőségek

- Az AAindex kb. 500 skálát tartalmaz – nem kell mind
- A környezet mekkora részét vegyük figyelembe? – ablakméret
- Milyen súllyal vegyük figyelembe a környezetet?
- Milyen módon vegyük figyelembe?
- Nem reménytelen ügy, de jobb lesz-e így?

Mit nyerhetünk egy új modellel

C-F

- 'szerkezet képző hajlam' – 1D
- Csúszóablakos összegzés
- A paraméterek optimalizása kísérletes eredmények alapján

SVM

- Vektorleírás – ND
- Szekvenciális környezet figyelembe vétele
- Súlyozás
- A modell tanítása kísérletes eredmények alapján

Mit tanultunk ma?

- A statisztikus modellek:
 - nagyon elterjedtek a bioinformatikában
 - többnyire egyszerű elvekre épülnek
 - tartalmaznak belső paramétereket
 - ezeket kalibrálni kell
- Végő soron a kísérletes adatok minőségén múlik a modell jósága



Feladat 7.

- A nevedből képzett fehérje szekvencia C-F predikció alapján inkább helix vagy redő szerkezetű-e?